

# From predictive to interactive multimodal language learning

by

Angeliki Lazaridou

Submitted to the Center for Mind and Brain Sciences (CiMeC)  
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

UNIVERSITY OF TRENTO

December 2016

© University of Trento 2016. All rights reserved.

Author .....  
Center for Mind and Brain Sciences (CiMeC)  
December 2, 2016

Certified by .....  
Marco Baroni  
Associate Professor  
Thesis Supervisor



# From predictive to interactive multimodal language learning

by

Angeliki Lazaridou

Submitted to the Center for Mind and Brain Sciences (CiMeC)  
on December 2, 2016, in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

## Abstract

The way humans learn the meaning of words is a fundamental question in many different disciplines and, from a computational perspective, an answer to this question could lead to important advances in artificial intelligence. While the details of the learning process are still an open question, what we do know is that humans make use of the very rich perceptual input present in the communicative setups in which learning takes place.

In this work, we will present three models of human learning from naturalistic multi-modal input. We will start by introducing a model that assumes a purely predictive learner existing in a non-communicative setup and show that such a computational learner when tested displays comparable learning behaviour to human learners on a novel word learning setup. We will then relax some of the learning assumptions and present a model that, instead of exposing the computational learner to a passive environment, such as the text corpora traditionally used in semantic learning experiments, it exposes the learner to communicative episodes, simulated in our experiments by corpora capturing multi-modal interactions between children and their caregivers, allowing the learner to make use of information beyond words and passive percept during learning. Finally, we will present on-going work towards interactive learning between two agents.

Thesis Supervisor: Marco Baroni

Title: Associate Professor





# Acknowledgments

They say “*It’s not the destination, but the journey*”; simply put, these 4 years have been the best years of my life!!!!!!!!!!!!!!

Marco Baroni, my thesis advisor, is of course responsible for a couple (at least) of those exclamation marks. He has been an incredible teacher, shaping but never imposing ideas leaving me space to develop as a researcher, an insightful advisor, giving always the right thought-provoking feedback that can turn a half-baked, crazy idea into something concrete, an amazing collaborator, putting tons amount of work on our papers, a positive person for bringing some much good energy in our meetings and always seeing the glass half-full, and one h\*\*l of a very patient person for being able to cope with me all these years. I’ve been truly fortunate for having worked with him – I will consider myself lucky to become one day 1/10 of the researcher he is....

Not only is Marco an inherently cool person, but he managed the impossible, i.e., to put together a dream team of researchers and PhD students, the **COMPOSES** group. Nghia The Pham, Georgiana Dinu and German Kruszewski my super smart colleagues in the office and best friends in life, who supported me in every single step of this PhD, who taught me an insane amount of things, who were always eager to feedback on whatever crazy I was working on and with whom I had so much fun in conferences, Stapomatto aperitivos, in the office, outside of the office etc. Special thanks to Marco Marelli, with whom not only I had so many amusing afternoons in the aforementioned Stapomatto aperitivos, but also very creative conversations that lead in us co-authoring the material presented in Chapter 3.

Raffaella Bernardi, Denis Paperno, Roberto Zamparelli, Gemma Boleda, Aurelie Herbelot, Elia Bruni and Eva Maria Vecchi have helped me in different phases of these 4 years, either by providing moral support, feedback in my work, financial contribution for conferences, fun moments in the office to cope with stressful deadlines etc. I came to feel COMPOSES not just as an academic group, but as a family. I really hope that my future working places will manage to live up to the standards (scientific and more) that COMPOSES has created.

My mother, grandmother and grandfather had always believed in me, letting me explore whatever was there in the world to explore, always supporting me in every possible way. I owe them eternal gratitude for the person I am today.

Finally, I would like to thank Adam. He has been always the first to hear my work-related ideas happy to provide very critical feedback, helping me when needing help, supporting me when needing support, making me laugh when wanting to cry, preparing food when I was hungry while at the same time I was running out of time for some of my endless deadlines. This world is just infinitely better when he is around...

The research was supported by a European Research Council Grant nr. 283554 (COMPOSES project).

This doctoral thesis has been examined by a Committee as follows:

Professor Marco Baroni .....  
Thesis Advisor

Professor Louise McNally .....  
Thesis Committee

Professor Hinrich Schütze .....  
Thesis Committee

Professor Roberto Zamparelli .....  
Thesis Committee

# Publications

Parts of this thesis (ideas, figures, results and discussions) have appeared previously in the following publications:

- Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skip-gram model. In *Proceedings of NAACL*, pages 153–163, Denver, CO, 2015.
- Angeliki Lazaridou, Grzegorz Chrupala, Raquel Fernandez and Marco Baroni. Multimodal semantic learning from child-directed input. In *Proceedings of NAACL*, pages 387–392, San Diego, CA, 2016.
- Angeliki Lazaridou, Marco Marelli, and Marco Baroni. Multimodal word meaning induction from minimal exposure to natural text. In *Cognitive Science*, To appear.
- Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Towards Multi-Agent Communication-Based Language Learning. arXiv e-print 1605.07133.

# Contents

<b>1</b>	<b>Introduction</b>	<b>19</b>
<b>2</b>	<b>Multimodal Skip-gram Model</b>	<b>25</b>
2.1	Introduction . . . . .	25
2.2	Related Work . . . . .	26
2.3	Multimodal Skip-gram Architecture . . . . .	28
2.3.1	Skip-gram Model . . . . .	28
2.3.2	Injecting visual knowledge . . . . .	29
2.3.3	Multi-modal Skip-gram Model A MSG-A . . . . .	30
2.3.4	Multi-modal Skip-gram Model B MSG-B . . . . .	31
2.4	Experimental Setup . . . . .	31
2.5	Experiments . . . . .	32
2.5.1	Approximating human judgments . . . . .	32
2.5.2	Zero-shot image labeling and retrieval . . . . .	35
2.5.3	Abstract words . . . . .	37
2.6	Discussion . . . . .	41
<b>3</b>	<b>Multimodal word meaning induction from minimal exposure to natural text</b>	<b>43</b>
3.1	Introduction . . . . .	43
3.2	Experimental setup . . . . .	48
3.2.1	Experimental materials . . . . .	48
3.2.2	Computational model . . . . .	55

3.3	Experiment 1: Estimating relatedness of chimeric concepts to word probes . . . . .	57
3.3.1	Methods . . . . .	57
3.3.2	Results . . . . .	60
3.4	Experiment 2: Estimating relatedness of chimeric concepts to image probes . . . . .	63
3.4.1	Methods . . . . .	63
3.4.2	Results . . . . .	65
3.5	Discussion . . . . .	68
<b>4</b>	<b>Attentive Social Multimodal Skip-gram Model</b>	<b>73</b>
4.1	Introduction . . . . .	73
4.2	Related Work . . . . .	74
4.3	Attentive Social MSG Model . . . . .	76
4.4	The Illustrated Frank et al. Corpus . . . . .	77
4.5	Experiments . . . . .	79
4.6	Discussion . . . . .	82
<b>5</b>	<b>Towards Multi-Agent Communication-Based Language Learning</b>	<b>85</b>
5.1	Introduction . . . . .	85
5.2	A two-agent referential game simulation . . . . .	88
5.2.1	The game . . . . .	88
5.2.2	Visual Scenes . . . . .	89
5.2.3	Agent Players . . . . .	91
5.3	Experiments . . . . .	94
5.3.1	General Training Details . . . . .	94
5.3.2	Results . . . . .	95
5.4	Discussion . . . . .	99
<b>6</b>	<b>Conclusion</b>	<b>101</b>

# List of Figures

2-1	“Cartoon” of MMSKIP-GRAM-B. Linguistic context vectors are actually associated to classes of words in a tree, not single words. SKIP-GRAM is obtained by ignoring the visual objective, MMSKIP-GRAM-A by fixing $M^{u \rightarrow v}$ to the identity matrix. . . . .	29
2-2	Examples of nearest visual neighbours of some abstract words: on the left, cases where subjects preferred the neighbour to the random foil; on the right, cases where they did not. . . . .	39
2-3	(a) Distribution of MSG-A vector activation for <i>meat</i> (blue) and <i>hope</i> (red). (b) Spearman $\rho$ between concreteness and various measures on the [50] set. . . . .	41
3-1	Example of the materials used in an experimental trial for the <i>gorilla/bear</i> chimera, corresponding to the nonce string MOHALK (the original word in each sentence, shown in squared brackets, was not presented to subjects). Unrelated and related probes are shown bottom left and right, respectively (in Experiment 1, only word probes were presented, in Experiment 2, only images). . . . .	53
3-2	Example of the materials used in an experimental trial for the <i>gorilla/bear</i> chimera, corresponding to the nonce string MOHALK (the original word in each sentence, shown in squared brackets, was not presented to subjects). Unrelated and related probes are shown bottom left and right, respectively (in Experiment 1, only word probes were presented, in Experiment 2, only images). . . . .	58

3-3	Experiment 1 (word probes): association between by-item average human responses and model predictions. Yellow circles: 2-sentence passages; orange triangles: 4-sentence passage; red squares: 6-sentence passages. . . . .	62
3-4	Experiment 1 (word probes): association between CPR and by-item average human responses (left-hand panel) and between CPR and model predictions (right-hand panel). Yellow circles: 2-sentence passages; orange triangles: 4-sentence passage; red squares: 6-sentence passages.	63
3-5	Experiment 1 (word probes): Interaction between informativeness and CPR for human responses (left panel) and model predictions (right panel). . . . .	64
3-6	Experiment 2 (image probes): association between by-item average human responses and model predictions. Yellow circles: 2-sentence passages; orange triangles: 4-sentence passage; red squares: 6-sentence passages. . . . .	66
3-7	Experiment 2 (image probes): association between CPR and by-item average human responses (left-hand panel) and between CPR and model predictions (right-hand panel). Yellow circles: 2-sentence passages; orange triangles: 4-sentence passage; red squares: 6-sentence passages. . . . .	67
3-8	Experiment 2 (image probes): Interaction between informativeness and CPR on human responses (left panel) and model predictions (right panel).	69
3-9	How MSG visualizes the novel word in each of 4 passages, two constructed from <i>caterpillar/cockroach</i> contexts, and two from <i>potato/turnip</i> contexts. . . . .	70
4-1	Fragment of the IFC corpus where symbolic labels <b>ring</b> and <b>hat</b> have been replaced by real images. Red frames mark objects being touched by the caregiver. . . . .	77



5-1	Schematic representation of the referential game. Agent A1 has to describe the dashed object with an attribute. Agent A2, has to guess which object A1 was referring to. In this example, the agents agree on the referent, so the communicative act was successful. . . . .	87
5-2	Neural network of player A1. In this particular game, A1 produces the reference vector that activates <code>word_2</code> as the RE, which is going to be passed over to A2. . . . .	92
5-3	Network of A2. In this example, given the two objects and the reference vector encoding the predicted word produced by A1, she correctly predicts that the referent is the second object. . . . .	93
5-4	Communication success across different datasets over training epochs. For each dataset, we vary the number of words available in the vocabulary of the agents, with the corresponding curves depicted in different colors in the figure <i>Note: Best viewed in color.</i> . . . . .	94
5-5	Inferred alignment of gold terms to the induced ones in the <b>Shapes dataset</b> . . . . .	97
5-6	Pairwise cosine similarities of gold terms in induced-word vector space in the <b>Objects</b> dataset when ordered randomly ( <b>up</b> ) and according to their category ( <b>down</b> ). . . . .	99



# List of Tables

2.1	Spearman correlation between model-generated similarities and human judgments. Right columns report correlation on visual-coverage subsets (percentage of original benchmark covered by subsets on first row of respective columns). First block reports results for out-of-the-box models; second block for visual and textual representations alone; third block for our implementation of multimodal models. . . . .	33
2.2	Ordered top 3 neighbours of example words in purely textual and multimodal spaces. Only <i>donut</i> and <i>owl</i> were trained with direct visual information. . . . .	34
2.3	Percentage precision@ $k$ results in the zero-shot image labeling task. .	37
2.4	Percentage precision@ $k$ results in the zero-shot image retrieval task. .	37
2.5	Subjects' preference for nearest visual neighbour of words in Kiela et al. (2014) vs. random pictures. Figure of merit is percentage proportion of significant results in favor of nearest neighbour across words. Results are reported for the whole set, as well as for words above ( <i>concrete</i> ) and below ( <i>abstract</i> ) the concreteness rating median. The <i>unseen</i> column reports results when words exposed to direct visual evidence during training are discarded. The <i>words</i> columns report set cardinality. . .	39

3.1	Chimeras and probes. For each chimera, the table reports, in this order, the chimera pivot and compatible components, their McRae similarity (and rank of compatible item among neighbours of pivot) and the probes with (left) word-based and (right) image-based CPR scores. . . . .	52
3.2	Rating distributions in Experiment 2. Average ratings (in different columns) are reported for set of probes grouped and ranked by their ground-truth CPR, crossed with passage informativeness (upper table) and passage length (lower table). . . . .	61
3.3	Experiment 1 (word probes): Results when analyzing either human responses or model predictions. . . . .	61
3.4	Rating distributions in Experiment 2. Average ratings (in different columns) are reported for set of probes grouped and ranked by their ground-truth CPR, crossed with passage informativeness (upper table) and passage length (lower table). . . . .	65
3.5	Experiment 2 (image probes): Results when analyzing either human responses or model predictions. . . . .	68
4.1	Best-F results for the MSG variations and alternative models on word-object matching. For all MSG models, we report Best-F mean and standard deviation over 100 iterations. . . . .	79
4.2	Test words occurring only once in IFC, together with corresponding gold objects, AttentiveSocialMSG top visual neighbours among the test items and in a larger 5.1K-objects set, and ranks of gold object in the two confusion sets. . . . .	82
4.3	Generalization experiment. Search space of each concept containing 40 elements. 10 gold, 10 visually similar, 10 less similar, 10 random . . . . .	83
5.1	Examples of $\langle referent, context \rangle$ pairs from our 3 datasets. Referents are marked with a green square. . . . .	89

5.2	Proportion of referentially inconsistent words, i.e., $RI(a) > 0$ , across different datasets and with different word vocabulary size $ V $ . Smaller values reflect consistent use. . . . .	98
-----	--	----



# Chapter 1

## Introduction

How do humans learn and use language? This is fundamental question addressed by many different disciplines such as cognitive science, natural language processing but also machine learning, and from a computational perspective, an answer to this question could lead to important advances in artificial intelligence. Traditionally, computational models of meaning relying on corpus-extracted context vectors, such as LSA [57], HAL [62], Topic Models [37] and more recent neural-network approaches [20, 70] have successfully tackled a number of lexical semantics tasks, where context vector similarity highly correlates with various indices of semantic relatedness [107]. Given that these models are learned from naturally occurring data using simple associative techniques, various authors have advanced the claim that they might be also capturing some crucial aspects of how humans acquire and use language [57, 61].

However, the models induce the meaning of words entirely from their co-occurrence with other words, without links to the external world. This constitutes a serious blow to claims of cognitive plausibility in at least two respects. One is the *grounding problem* [38, 87]. Irrespective of their relatively high performance on various semantic tasks, it is debatable whether models that have no access to visual and perceptual information can capture the holistic, grounded knowledge that humans have about concepts. Even from an empirical perspective, there is increasing evidence in the literature in favor of grounded cognition, suggesting that humans acquire and use language within a multi-modal environment. In one of the classic studies of *embodied*

*cognition*, [78] found that when participants simply *read* the word for an action, the *motor system* became active to represent its meaning. In another study, [77] replaced *words* with *pictures*, and found that this did not disrupt sentence processing, suggesting that the pictures were integrated into multi-modal representations of sentence meaning. [36] showed that *gestures* can convey an emerging conceptualization that cannot yet be articulated in *speech*, thus complementing spoken language during communication. In [39] it was reported that *sentence processing* was faster when participants' faces were configured discretely into states associated with particular emotions and when *facial emotion* matched sentence emotion (see [10] for a complete review).

From both theoretical and empirical perspective it becomes clear that the computational models of meaning should have access to extra-linguistic information during learning. One possible source of such extra-linguistic information used by several studies on multi-modal learning [2, 89] is offered by subject-derived feature norms. Feature norms are procured by human participants and aiming at capturing knowledge about concepts that is acquired by making use of our rich sensori-motor system (e.g., tigers *has stripes*, lemons *taste bitter*, elephants *are big* etc.). While this is a great proxy to our conceptual knowledge, feature norms are only available for a limited number of concrete concepts, preventing large-scale multi-modal learning. Moreover, since feature norms are the output of humans rather than naturalistic input, they are not suited for conducting realistic simulations of multi-modal learning. As a way to circumvent the short-comings of feature norms, very recently, a number of papers have exploited advances in automated feature extraction from images and videos to conduct multi-modal learning [16, 28, 49, 88].

Multi-modal learning, while a necessary condition for having human-like machines to which we can talk, is hardly a sufficient one. One important aspect of language is its interactive nature, the fact that we use language to communicate. This aspect however is heavily neglected in the current mainstream approach to train natural language systems, which instead expose machines to large amounts of text and having them perform passive learning (e.g., as in cases of image captioning or machine



translation [101, 117]). However, while this paradigm is an excellent way to learn general statistical associations between sequences of symbols, it focuses on the structure of language rather than on its purpose, that is, the fact that we use words to make things happen [3]!

In this work, we will present three models of human learning from naturalistic multi-modal input. We will start by introducing a model that assumes a purely predictive learner existing in a non-communicative setup and show that such a computational learner displays comparable learning behaviour as human learners on a novel word learning setup. We will then relax some of the learning assumptions and present a model that instead of placing the computational learner in a passive environment, such as corpora like Wikipedia traditionally used in semantic learning experiments, it places the learner in communicative episodes, simulated in our experiments by corpora capturing multi-modal interactions between children and their caregivers, allowing the learner to make use of information beyond words during learning. Finally, we will present on-going work towards interactive learning between two agents.

The thesis is structured as follows:

**Chapter 2** We introduce two new models for multi-modal learning, the *Multi-modal Skipgram A* and *Multi-modal Skipgram B* (MSG-A and MSG-B respectively), which build upon the very effective skip-gram approach of [71]. Unlike previous work on multi-modal models of semantic learning that assumes access to both linguistic and visual information for all words, our model, because its joint language-vision objective encourages the propagation of visual information to representations of words for which no direct visual evidence was used, thus allowing knowledge generalization across modalities. The resulting multimodally-enhanced vectors achieve remarkably good performance both on traditional semantic benchmarks, and in their new application to the “zero-shot” [31, 96] image labeling and retrieval scenario. Moreover, we show that indirect visual evidence also affects the representation of abstract words, paving the way to new cognitive studies and novel applications in computer vision.

**Chapter 3** We report a study on multimodal semantic learning from minimal exposure to natural text. In this study, we first run a set of experiments investigating whether minimal distributional evidence from very short passages extracted from realistic corpora suffices to trigger successful word learning in subjects, testing their linguistic and visual intuitions about the concepts associated to new words. After confirming that subjects are indeed very efficient distributional learners even from small amounts of evidence, we test the previously-introduced MSG-B model on the same multimodal task, finding that it behaves in a remarkable human-like way. We conclude that distributional semantic models provide a convincing computational account of word learning even at the early stages in which a word is first encountered, and that the way they build meaning representations can offer new insights into human language acquisition.

**Chapter 4** We take a step towards a more realistic setup by introducing a model that operates on naturalistic images of the objects present in a communicative episode. We build upon MSG-A and enhance it to handle cross-referential uncertainty. Moreover, we extend the cues commonly used in multimodal learning (e.g., objects in the environment) to include social cues (e.g., eyegaze, gestures, body posture, etc.) that reflect speakers’ intentions and generally contribute to the unfolding of the communicative situation [100]. Specifically we incorporate information regarding the objects that caregivers are holding. We show that our model, ATTENTIVE SOCIAL MSG, is able to integrate linguistic and extra-linguistic information (visual and social cues), handles referential uncertainty, and correctly learns to associate words with objects, even in cases of limited linguistic exposure.

**Chapter 5** We present a proposal for developing intelligent agents with language capabilities, that breaks away from current passive supervised regimes; agents co-exist and are able to interact with each other. In our framework, we consider the most basic act of communication, i.e., learning to refer to things, and we designed a “grounded” cooperative task that takes the form of referential games played by

two agents. The first experiments, while encouraging, revealed that it is essential to ensure that agents will not “drift” into their own language, but instead they will evolve one that is aligned to our natural languages.



# Chapter 2

## Multimodal Skip-gram Model

### 2.1 Introduction

Distributional semantic models (DSMs) derive vector-based representations of meaning from patterns of word co-occurrence in corpora. DSMs have been very effectively applied to a variety of semantic tasks [19, 71, 107]. However, compared to human semantic knowledge, these purely textual models, just like traditional symbolic AI systems [38, 87], are severely impoverished, suffering of *lack of grounding* in extra-linguistic modalities [35]. This observation has led to the development of *multimodal distributional semantic models* (MDSMs) [16, 28, 90], that enrich linguistic vectors with perceptual information, most often in the form of visual features automatically induced from image collections.

MDSMs outperform state-of-the-art text-based approaches, not only in tasks that directly require access to visual knowledge [15], but also on general semantic benchmarks [16, 90]. However, current MDSMs still have a number of drawbacks. First, they are generally constructed by first separately building linguistic and visual representations of the same concepts, and then merging them. This is obviously very different from how humans learn about concepts, by hearing words in a situated perceptual context. Second, MDSMs assume that both linguistic and visual information is available for all words, with no generalization of knowledge across modalities. Third, because of this latter assumption of full linguistic and visual coverage, current

MDSMs, paradoxically, cannot be applied to computer vision tasks such as image labeling or retrieval, since they do not generalize to images or words beyond their training set.

We introduce the *multimodal skip-gram* models, two new MDSMs that address all the issues above. The models build upon the very effective skip-gram approach of [70], that constructs vector representations by learning, incrementally, to predict the linguistic contexts in which target words occur in a corpus. In our extension, for a subset of the target words, relevant visual evidence from natural images is presented together with the corpus contexts (just like humans hear words accompanied by concurrent perceptual stimuli). The model must learn to predict these visual representations jointly with the linguistic features. The joint objective encourages the propagation of visual information to representations of words for which no direct visual evidence was available in training. The resulting multimodally-enhanced vectors achieve remarkably good performance both on traditional semantic benchmarks, and in their new application to the “zero-shot” image labeling and retrieval scenario. Very interestingly, indirect visual evidence also affects the representation of abstract words, paving the way to ground-breaking cognitive studies and novel applications in computer vision.

## 2.2 Related Work

There is by now a large literature on multimodal distributional semantic models. We focus here on a few representative systems. [16] propose a straightforward approach to MDSM induction, where text- and image-based vectors for the same words are constructed independently, and then “mixed” by applying the Singular Value Decomposition to their concatenation. An empirically superior model has been proposed by [90], who use more advanced visual representations relying on images annotated with high-level “visual attributes”, and a multimodal fusion strategy based on stacked autoencoders. [49] adopt instead a simple concatenation strategy, but obtain empirical improvements by using state-of-the-art convolutional neural networks to extract

visual features, and the skip-gram model for text. These and related systems take a two-stage approach to derive multimodal spaces (unimodal induction followed by fusion), and they are only tested on concepts for which both textual and visual labeled training data are available (the pioneering model of [28] did learn from text and images jointly using Topic Models, but was shown to be empirically weak by [16]).

[42] propose an incremental multimodal model based on simple recurrent networks [25], focusing on *grounding propagation* from early-acquired concrete words to a larger vocabulary. However, they use subject-generated features as surrogate for realistic perceptual information, and only test the model in small-scale simulations of word learning. [41], whose evaluation focuses on how perceptual information affects different word classes more or less effectively, similarly to Howell et al., integrate perceptual information in the form of subject-generated features and text from image annotations into a skip-gram model. They inject perceptual information by merging words expressing perceptual features with corpus contexts, which amounts to linguistic-context re-weighting, thus making it impossible to separate linguistic and perceptual aspects of the induced representation, and to extend the model with non-linguistic features. We use instead authentic image analysis as proxy to perceptual information, and we design a robust way to incorporate it, easily extendible to other signals, such as feature norm or brain signal vectors [32].

The recent work on so-called *zero-shot learning* to address the annotation bottleneck in image labeling [31, 58, 97] looks at image- and text-based vectors from a different perspective. Instead of combining visual and linguistic information in a common space, it aims at learning a mapping from image- to text-based vectors. The mapping, induced from annotated data, is then used to project images of objects that were not seen during training onto linguistic space, in order to retrieve the nearest word vectors as labels. Multimodal word vectors should be better-suited than purely text-based vectors for the task, as their similarity structure should be closer to that of images. However, traditional MDSMs cannot be used in this setting, because they do not cover words for which no manually annotated training images are available, thus defeating the generalizing purpose of zero-shot learning. We will show below

that our multimodal vectors, that are not hampered by this restriction, do indeed bring a significant improvement over purely text-based linguistic representations in the zero-shot setup.

Multimodal language-vision spaces have also been developed with the goal of improving caption generation/retrieval and caption-based image retrieval [45, 53, 64, 98]. These methods rely on necessarily limited collections of captioned images as sources of multimodal evidence, whereas we automatically enrich a very large corpus with images to induce general-purpose multimodal word representations, that could be used as input embeddings in systems specifically tuned to caption processing. Thus, our work is complementary to this line of research.

## 2.3 Multimodal Skip-gram Architecture

### 2.3.1 Skip-gram Model

We start by reviewing the standard SKIP-GRAM model of [70], in the version we use. Given a text corpus, SKIP-GRAM aims at inducing word representations that are good at predicting the *context* words surrounding a *target* word. Mathematically, it maximizes the objective function:

$$\frac{1}{T} \sum_{t=1}^T \left( \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j}|w_t) \right) \quad (2.1)$$

where  $w_1, w_2, \dots, w_T$  are words in the training corpus and  $c$  is the size of the window around target  $w_t$ , determining the set of context words to be predicted by the induced representation of  $w_t$ . Following Mikolov et al., we implement a subsampling option randomly discarding context words as an inverse function of their frequency, controlled by hyperparameter  $t$ . The probability  $p(w_{t+j}|w_t)$ , the core part of the objective in Equation 2.1, is given by softmax:

$$p(w_{t+j}|w_t) = \frac{e^{u'_{w_{t+j}}{}^T u_{w_t}}}{\sum_{w'=1}^W e^{u'_{w'}{}^T u_{w_t}}} \quad (2.2)$$



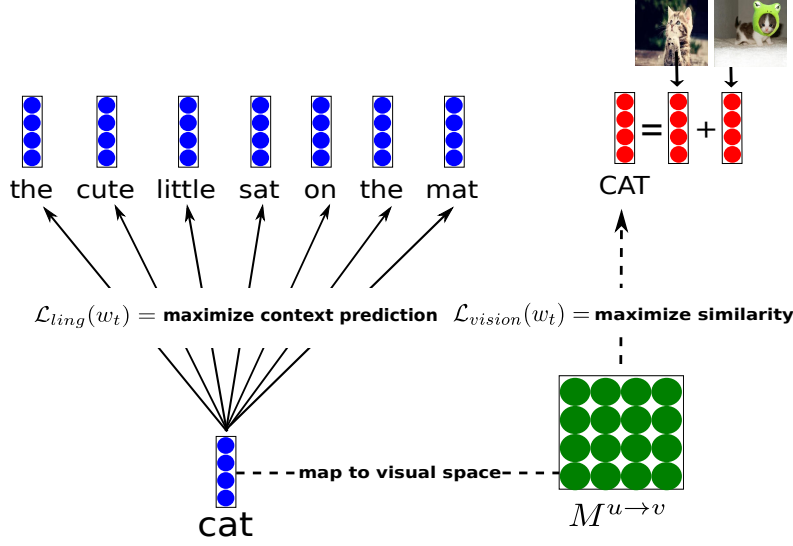


Figure 2-1: “Cartoon” of MMSKIP-GRAM-B. Linguistic context vectors are actually associated to classes of words in a tree, not single words. SKIP-GRAM is obtained by ignoring the visual objective, MMSKIP-GRAM-A by fixing  $M^{u \rightarrow v}$  to the identity matrix.

where  $u_w$  and  $u'_w$  are the context and target vector representations of word  $w$  respectively, and  $W$  is the size of the vocabulary. Due to the normalization term, Equation 2.2 requires  $O(|W|)$  time complexity. A considerable speedup to  $O(\log|W|)$ , is achieved by using the hierarchical version of Equation 2.2 [74], adopted here.

### 2.3.2 Injecting visual knowledge

We now assume that word learning takes place in a *situated* context, in which, for a subset of the target words, the corpus contexts are accompanied by a visual representation of the concepts they denote (just like in a conversation, where a linguistic utterance will often be produced in a visual scene including some of the word referents). The visual representation is also encoded in a vector (we describe in Section 3.2 below how we construct it). We thus make the skip-gram “multimodal” by adding a second, *visual* term to the original *linguistic* objective, that is, we extend Equation 2.1 as follow:

$$\frac{1}{T} \sum_{t=1}^T (\mathcal{L}_{ling}(w_t) + \mathcal{L}_{vision}(w_t)) \quad (2.3)$$

where  $\mathcal{L}_{ling}(w_t)$  is the text-based skip-gram objective  $\sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j}|w_t)$ , whereas the  $\mathcal{L}_{vision}(w_t)$  term forces word representations to take visual information into account. Note that if a word  $w_t$  is not associated to visual information, as is systematically the case, e.g., for determiners and non-imageable nouns, but also more generally for any word for which no visual data are available,  $\mathcal{L}_{vision}(w_t)$  is set to 0.

We now propose two variants of the visual objective, resulting in two distinguished multi-modal versions of the skip-gram model.

### 2.3.3 Multi-modal Skip-gram Model A MSG-A

One way to force word embeddings to take visual representations into account is to try to directly increase the similarity (expressed, for example, by the *cosine*) between linguistic and visual representations, thus aligning the dimensions of the linguistic vector with those of the visual one (recall that we are inducing the first, while the second is fixed), and making the linguistic representation of a concept “move” closer to its visual representation. We maximize similarity through a max-margin framework commonly used in models connecting language and vision [113, 31]. More precisely, we formulate the visual objective  $\mathcal{L}_{vision}(w_t)$  as:

$$- \sum_{w' \sim P_n(w)} \max(0, \gamma - \cos(u_{w_t}, v_{w_t}) + \cos(u_{w_t}, v_{w'})) \quad (2.4)$$

where the minus sign turns a loss into a cost,  $\gamma$  is the margin,  $u_{w_t}$  is the target multimodally-enhanced word representation we aim to learn,  $v_{w_t}$  is the corresponding visual vector (fixed in advance) and  $v_{w'}$  ranges over visual representations of words (featured in our image dictionary) randomly sampled from distribution  $P_n(w_t)$ . These random visual representations act as “negative” samples, encouraging  $u_{w_t}$  to be more similar to its own visual representation than to that of other words. The sampling distribution is currently set to uniform, and the number of negative samples controlled by hyperparameter  $k$ .

### 2.3.4 Multi-modal Skip-gram Model B MSG-B

The visual objective in MSG-A has the drawback of assuming a direct comparison of linguistic and visual representations, constraining them to be of equal size. MSG-B lifts this constraint by including an extra layer mediating between linguistic and visual representations (see Figure 2-1 for a sketch of MSG-B). Learning this layer is equivalent to estimating a cross-modal mapping matrix from linguistic onto visual representations, jointly induced with linguistic word embeddings. The extension is straightforwardly implemented by substituting, into Equation 2.4, the word representation  $u_{w_t}$  with  $z_{w_t} = M^{u \rightarrow v} u_{w_t}$ , where  $M^{u \rightarrow v}$  is the cross-modal mapping matrix to be induced. To avoid overfitting, we also add an L2 regularization term for  $M^{u \rightarrow v}$  to the overall objective (Equation 2.3), with its relative importance controlled by hyperparameter  $\lambda$ .

## 2.4 Experimental Setup

The parameters of all models are estimated by back-propagation of error via stochastic gradient descent. Our text corpus is a Wikipedia 2009 dump comprising approximately 800M tokens.<sup>1</sup> To train the multimodal models, we add visual information for 5,100 words that have an entry in ImageNet [23], occur at least 500 times in the corpus and have concreteness score  $\geq 0.5$  according to [106]. On average, about 5% tokens in the text corpus are associated to a visual representation. To construct the visual representation of a word, we sample 100 pictures from its ImageNet entry, and extract a 4096-dimensional vector from each picture using the Caffe toolkit [43], together with the pre-trained convolutional neural network of [54]. The vector corresponds to activation in the top (FC7) layer of the network. Finally, we average the vectors of the 100 pictures associated to each word, deriving 5,100 aggregated visual representations.

---

<sup>1</sup><http://wacky.sslmit.unibo.it>

**Hyperparameters** For both SKIP-GRAM and the MSG models, we fix hidden layer size to 300. To facilitate comparison between MSG-A and MSG-B, and since the former requires equal linguistic and visual dimensionality, we keep the first 300 dimensions of the visual vectors. For the linguistic objective, we use hierarchical softmax with a Huffman frequency-based encoding tree, setting frequency subsampling option  $t=0.001$  and window size  $c=5$ , without tuning. The following hyperparameters were tuned on the text9 corpus:<sup>2</sup> MSG-A:  $k=20$ ,  $\gamma=0.5$ ; MSG-B:  $k=5$ ,  $\gamma=0.5$ ,  $\lambda=0.0001$ .

## 2.5 Experiments

### 2.5.1 Approximating human judgments

**Benchmarks** A widely adopted way to test DSMs and their multimodal extensions is to measure how well model-generated scores approximate human similarity judgments about pairs of words. We put together various benchmarks covering diverse aspects of meaning, to gain insights on the effect of perceptual information on different similarity facets. Specifically, we test on general relatedness (*MEN*, [16], 3K pairs), e.g., pickles are related to hamburgers, semantic ( $\approx$  taxonomic) similarity (*Simlex-999*, [40], 1K pairs; *SemSim*, [90], 7.5K pairs), e.g., pickles are similar to onions, as well as visual similarity (*VisSim*, [90], same pairs as SemSim with different human ratings), e.g., pickles look like zucchinis.

**Alternative Multimodal Models** We compare our models against several recent alternatives. We test the vectors made available by [49]. Similarly to us, they derive textual features with the skip-gram model (from a portion of the Wikipedia and the British National Corpus) and use visual representations extracted from the ESP dataset [110] through a convolutional neural network [75]. They concatenate textual and visual features after normalizing to unit length and centering to zero mean. We also test the vectors that performed best in the evaluation of [16], based on textual features extracted from a 3B-token corpus and SIFT-based Bag-of-Visual-Words visual

---

<sup>2</sup><http://mattmahoney.net/dc/textdata.html>

Model	<i>MEN</i>		<i>Simlex-999</i>		<i>SemSim</i>		<i>VisSim</i>	
	100%	42%	100%	29%	100%	85%	100%	85%
KIELA AND BOTTOU	-	0.74	-	0.33	-	0.60	-	0.50
BRUNI ET AL.	-	0.77	-	0.44	-	0.69	-	0.56
SILBERER AND LAPATA	-	-	-	-	0.70	-	0.64	-
CNN FEATURES	-	0.62	-	0.54	-	0.55	-	0.56
SKIP-GRAM	0.70	0.68	0.33	0.29	0.62	0.62	0.48	0.48
CONCATENATION	-	0.74	-	0.46	-	0.68	-	0.60
SVD	0.61	0.74	0.28	0.46	0.65	0.68	0.58	0.60
MSG-A	0.75	0.74	0.37	0.50	0.72	0.72	0.63	0.63
MSG-B	0.74	0.76	0.40	0.53	0.66	0.68	0.60	0.60

Table 2.1: Spearman correlation between model-generated similarities and human judgments. Right columns report correlation on visual-coverage subsets (percentage of original benchmark covered by subsets on first row of respective columns). First block reports results for out-of-the-box models; second block for visual and textual representations alone; third block for our implementation of multimodal models.

features [93] extracted from the ESP collection. Bruni and colleagues fuse a weighted concatenation of the two components through SVD. We further re-implement both methods with our own textual and visual embeddings as CONCATENATION and SVD (with target dimensionality 300, picked without tuning). Finally, we present for comparison the results on *SemSim* and *VisSim* reported by [90], obtained with a stacked-autoencoders architecture run on textual features extracted from Wikipedia with the Strudel algorithm [7] and attribute-based visual features [26] extracted from ImageNet.

All benchmarks contain a fair amount of words for which we did not use direct visual evidence. We are interested in assessing the models both in terms of how they fuse linguistic and visual evidence when they are both available, and for their robustness in lack of full visual coverage. We thus evaluate them in two settings. The visual-coverage columns of Table 2.1 (those on the right) report results on the subsets for which all compared models have access to direct visual information for both words. We further report results on the full sets (“100%” columns of Table 2.1) for models that can propagate visual information and that, consequently, can meaningfully be tested on words without direct visual representations.

<i>Target</i>	SKIP-GRAM	MSG-A	MSG-B
donut	fridge, diner, candy	pizza, sushi, sandwich	pizza, sushi, sandwich
owl	pheasant, woodpecker, squirrel	eagle, woodpecker, falcon	eagle, falcon, hawk
mural	sculpture, painting, portrait	painting, portrait, sculpture	painting, portrait, sculpture
tobacco	coffee, cigarette, corn	cigarette, cigar, corn	cigarette, cigar, smoking
depth	size, bottom, meter	sea, underwater, level	sea, size, underwater
chaos	anarchy, despair, demon	demon, anarchy, destruction	demon, anarchy, shadow

Table 2.2: Ordered top 3 neighbours of example words in purely textual and multi-modal spaces. Only *donut* and *owl* were trained with direct visual information.

**Results** The state-of-the-art visual CNN FEATURES alone perform remarkably well, outperforming the purely textual model (SKIP-GRAM) in two tasks, and achieving the best absolute performance on the visual-coverage subset of Simlex-999. Regarding multimodal *fusion* (that is, focusing on the visual-coverage subsets), both MMSKIP-GRAM models perform very well, at the top or just below it on all tasks, with comparable results for the two variants. Their performance is also good on the full data sets, where they consistently outperform SKIP-GRAM and SVD (that is much more strongly affected by lack of complete visual information). They’re just a few points below the state-of-the-art MEN correlation (0.8), achieved by [8] with a corpus 3 larger than ours and extensive tuning. MSG-B is close to the state of the art for Simlex-999, reported by the resource creators to be at 0.41 [40]. Most impressively, MSG-A reaches the performance level of the [90] model on their SemSim and VisSim data sets, despite the fact that the latter has full visual-data coverage and uses attribute-based image representations, requiring supervised learning of attribute classifiers, that achieve performance in the semantic tasks comparable or higher than that of our CNN features (see Table 3 in [90]). Finally, if the multimodal models (unsurprisingly) bring about a large performance gain over the purely linguistic model on visual similarity, the improvement is consistently large also for the other benchmarks, confirming that multimodality leads to better semantic models in general, that can help in capturing different types of similarity (general relatedness, strictly taxonomic, perceptual).

While we defer to further work a better understanding of the relation between multimodal grounding and different similarity relations, Table 2.2 provides qualita-

tive insights on how injecting visual information changes the structure of semantic space. The top SKIP-GRAM neighbours of *donuts* are places where you might encounter them, whereas the multimodal models relate them to other take-away food, ranking visually-similar pizzas at the top. The *owl* example shows how multimodal models pick taxonomically closer neighbours of concrete objects, since often closely related things also look similar [16]. In particular, both multimodal models get rid of squirrels and offer other birds of prey as nearest neighbours. No direct visual evidence was used to induce the embeddings of the remaining words in the table, that are thus influenced by vision only by propagation. The subtler but systematic changes we observe in such cases suggest that this indirect propagation is not only non-damaging with respect to purely linguistic representations, but actually beneficial. For the concrete *mural* concept, both multimodal models rank paintings and portraits above less closely related sculptures (they are not a form of painting). For *tobacco*, both models rank cigarettes and cigar over coffee, and MSG-B avoids the arguably less common “crop” sense cued by corn. The last two examples show how the multimodal models turn up the embodiment level in their representation of abstract words. For *depth*, their neighbours suggest a concrete marine setup over the more abstract measurement sense picked by the MSG neighbours. For *chaos*, they rank a demon, that is, a concrete agent of chaos at the top, and replace the more abstract notion of despair with equally gloomy but more imageable shadows and destruction (more on abstract words below).

### 2.5.2 Zero-shot image labeling and retrieval

The multimodal representations induced by our models should be better suited than purely text-based vectors to label or retrieve images. In particular, given that the quantitative and qualitative results collected so far suggest that the models propagate visual information across words, we apply them to image labeling and retrieval in the challenging zero-shot setup (see Section 2.2 above).<sup>3</sup>

---

<sup>3</sup>We will refer here, for conciseness’ sake, to *image* labeling/retrieval, but, as our visual vectors are aggregated representations of images, the tasks we’re modeling consist, more precisely, in labeling

**Setup** We take out as test set 25% of the 5.1K words we have visual vectors for. The multimodal models are re-trained without visual vectors for these words, using the same hyperparameters as above. For both tasks, the search for the correct word label/image is conducted on the whole set of 5.1K word/visual vectors.

In the image labeling task, given a visual vector representing an image, we map it onto word space, and label the image with the word corresponding to the nearest vector. To perform the vision-to-language mapping, we train a Ridge regression by 5-fold cross-validation on the test set (for SKIP-GRAM only, we also add the remaining 75% of word-image vector pairs used in estimating the multimodal models to the Ridge training data).<sup>4</sup>

In the image retrieval task, given a linguistic/multimodal vector, we map it onto visual space, and retrieve the nearest image. For SKIP-GRAM, we use Ridge regression with the same training regime as for the labeling task. For the multimodal models, since maximizing similarity to visual representations is already part of their training objective, we do not fit an extra mapping function. For MSG-A, we directly look for nearest neighbours of the learned embeddings in visual space. For MSG-B, we use the  $M^{u \rightarrow v}$  mapping function induced while learning word embeddings.

**Results** In image labeling (Table 2.3) SKIP-GRAM is outperformed by both multimodal models, confirming that these models produce vectors that are directly applicable to vision tasks thanks to visual propagation. The most interesting results however are achieved in image retrieval (Table 2.4), which is essentially the task the multimodal models have been implicitly optimized for, so that they could be applied to it without any specific training. The strategy of directly querying for the nearest visual vectors of the MSG-A word embeddings works remarkably well, outperforming on the higher ranks SKIP-GRAM, which requires an ad-hoc mapping function. This suggests that the multimodal embeddings we are inducing, while general enough to

---

a set of pictures denoting the same object and retrieving the corresponding set given the name of the object.

<sup>4</sup>We use one fold to tune Ridge  $\lambda$ , three to estimate the mapping matrix and test in the last fold. To enforce strict zero-shot conditions, we exclude from the test fold labels occurring in the LSVRC2012 set that was employed to train the CNN of [54], that we use to extract visual features.



	P@1	P@2	P@10	P@20	P@50
SKIP-GRAM	1.5	2.6	14.2	23.5	36.1
MSG-A	2.1	3.7	16.7	24.6	37.6
MSG-B	2.2	5.1	20.2	28.5	43.5

Table 2.3: Percentage precision@ $k$  results in the zero-shot image labeling task.

	P@1	P@2	P@10	P@20	P@50
SKIP-GRAM	1.9	3.3	11.5	18.5	30.4
MSG-A	1.9	3.2	13.9	20.2	33.6
MSG-B	1.9	3.8	13.2	22.5	38.3

Table 2.4: Percentage precision@ $k$  results in the zero-shot image retrieval task.

achieve good performance in the semantic tasks discussed above, encode sufficient visual information for direct application to image analysis tasks. This is especially remarkable because the word vectors we are testing were not matched with visual representations at model training time, and are thus multimodal only by propagation. The best performance is achieved by MSG-B, confirming our claim that its  $M^{u \rightarrow v}$  matrix acts as a multimodal mapping function.

### 2.5.3 Abstract words

We have already seen, through the *depth* and *chaos* examples of Table 2.2, that the indirect influence of visual information has interesting effects on the representation of abstract terms. The latter have received little attention in multimodal semantics, with [41] concluding that abstract nouns, in particular, do not benefit from propagated perceptual information, and their representation is even harmed when such information is forced on them (see Figure 4 of their paper). Still, embodied theories of cognition have provided considerable evidence that abstract concepts are also grounded in the senses [10, 56]. Since the word representations produced by MSG-A, including those pertaining to abstract concepts, can be directly used to search for near images in visual space, we decided to verify, experimentally, if these near images (of concrete things) are relevant not only for concrete words, as expected, but also

for abstract ones, as predicted by embodied views of meaning.

More precisely, we focused on the set of 200 words that were sampled across the USF norms concreteness spectrum by [50] (2 words had to be excluded for technical reasons). This set includes not only concrete (*meat*) and abstract (*thought*) nouns, but also adjectives (*boring*), verbs (*teach*), and even grammatical terms (*how*). Some words in the set have relatively high concreteness ratings, but are not particularly imageable, e.g.: *hot*, *smell*, *pain*, *sweet*. For each word in the set, we extracted the nearest neighbour picture of its MSG-A representation, and matched it with a random picture. The pictures were selected from a set of 5,100, all labeled with distinct words (the picture set includes, for each of the words associated to visual information as described in Section 3.2, the nearest picture to its aggregated visual representation). Since it is much more common for concrete than abstract words to be directly represented by an image in the picture set, when searching for the nearest neighbour we excluded the picture labeled with the word of interest, if present (e.g., we excluded the picture labeled *tree* when picking the nearest neighbour of the word *tree*). We ran a CrowdFlower<sup>5</sup> survey in which we presented each test word with the two associated images (randomizing presentation order of nearest and random picture), and asked subjects which of the two pictures they found more closely related to the word. We collected minimally 20 judgments per word. Subjects showed large agreement (median proportion of majority choice at 90%), confirming that they understood the task and behaved consistently.

We quantify performance in terms of proportion of words for which the number of votes for the nearest neighbour picture is significantly above chance according to a two-tailed binomial test. We set significance at  $p < 0.05$  after adjusting all p-values with the Holm correction for running 198 statistical tests. The results in Table 2.5 indicate that, in about half the cases, the nearest picture to a word MSG-A representation is meaningfully related to the word. As expected, this is more often the case for concrete than abstract words. Still, we also observe a significant preference for the model-predicted nearest picture for about one fourth of the abstract terms. Whether

---

<sup>5</sup><http://www.crowdflower.com>

	<i>global</i>	<i> words </i>	<i>unseen</i>	<i> words </i>
all	48%	198	30%	127
concrete	73%	99	53%	30
abstract	23%	99	23%	97

Table 2.5: Subjects’ preference for nearest visual neighbour of words in Kiela et al. (2014) vs. random pictures. Figure of merit is percentage proportion of significant results in favor of nearest neighbour across words. Results are reported for the whole set, as well as for words above (*concrete*) and below (*abstract*) the concreteness rating median. The *unseen* column reports results when words exposed to direct visual evidence during training are discarded. The *words* columns report set cardinality.



Figure 2-2: Examples of nearest visual neighbours of some abstract words: on the left, cases where subjects preferred the neighbour to the random foil; on the right, cases where they did not.

a word was exposed to direct visual evidence during training is of course making a big difference, and this factor interacts with concreteness, as only two abstract words were matched with images during training.<sup>6</sup> When we limit evaluation to word representations that were not exposed to pictures during training, the difference between concrete and abstract terms, while still large, becomes less dramatic than if all words are considered.

Figure 4-1 shows four cases in which subjects expressed a strong preference for the nearest visual neighbour of a word. *Freedom*, *god* and *theory* are strikingly in agreement with the view, from embodied theories, that abstract words are grounded in relevant concrete scenes and situations. The *together* example illustrates how visual

<sup>6</sup>In both cases, the images actually depict concrete senses of the words: a memory board for *memory* and a stop sign for *stop*.

data might ground abstract notions in surprising ways. For all these cases, we can borrow what [42] say about visual propagation to abstract words (p. 260):

Intuitively, this is something like trying to explain an abstract concept like *love* to a child by using concrete examples of scenes or situations that are associated with love. The abstract concept is never fully grounded in external reality, but it does inherit some meaning from the more concrete concepts to which it is related.

Of course, not all examples are good: the last column of Figure 4-1 shows cases with no obvious relation between words and visual neighbours (subjects preferred the random images by a large margin).

The multimodal vectors we induce also display an interesting intrinsic property related to the hypothesis that grounded representations of abstract words are more complex than for concrete ones, since abstract concepts relate to varied and composite situations [11]. A natural corollary of this idea is that visually-grounded representations of abstract concepts should be more diverse: If you think of dogs, very similar images of specific dogs will come to mind. You can also imagine the abstract notion of freedom, but the nature of the related imagery will be much more varied. Recently, [50] have proposed to measure abstractness by exploiting this very same intuition. However, they rely on manual annotation of pictures via Google Images and define an *ad-hoc* measure of *image dispersion*. We conjecture that the representations naturally induced by our models display a similar property. In particular, the *entropy* of our multimodal vectors, being an expression of how varied the information they encode is, should correlate with the degree of abstractness of the corresponding words. As Figure 2-3(a) shows, there is indeed a difference in entropy between the most concrete (*meat*) and most abstract (*hope*) words in the Kiela et al. set.

To test the hypothesis quantitatively, we measure the correlation of entropy and concreteness on the 200 words in the [50] set.<sup>7</sup> Figure 2-3(b) shows that the entropies of both the MSG-A representations and those generated by mapping MSG-B vectors onto visual space (MSG-B\*) achieve very high correlation (but, interestingly,

---

<sup>7</sup>Since the vector dimensions range over the real number line, we calculate entropy on vectors that are unit-normed after adding a small constant insuring all values are positive.

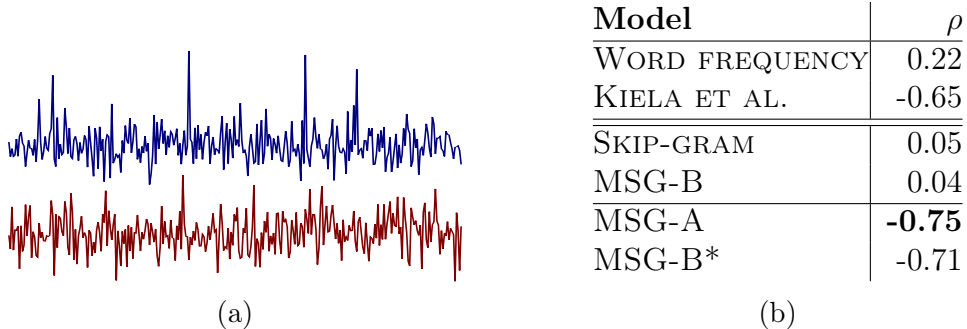


Figure 2-3: (a) Distribution of MSG-A vector activation for *meat* (blue) and *hope* (red). (b) Spearman  $\rho$  between concreteness and various measures on the [50] set.

not MSG-B). This is further evidence that multimodal learning is grounding the representations of both concrete and abstract words in meaningful ways.

## 2.6 Discussion

We introduced two multimodal extensions of SKIP-GRAM. MSG-A is trained by directly optimizing the similarity of words with their visual representations, thus forcing maximum interaction between the two modalities. MSG-B includes an extra mediating layer, acting as a cross-modal mapping component. The ability of the models to *integrate* and *propagate* visual information resulted in word representations that performed well in both semantic and vision tasks, and could be used as input in systems benefiting from prior visual knowledge (e.g., caption generation). Our results with abstract words suggest the models might also help in tasks such as metaphor detection, or even retrieving/generating pictures of abstract concepts. Their incremental nature makes them well-suited for cognitive simulations of grounded language acquisition, an avenue we are exploring in the next chapter.



# Chapter 3

## Multimodal word meaning induction from minimal exposure to natural text

### 3.1 Introduction

In the previous chapter we presented MSG, a computational learner able to learn word meanings from both linguistic and extra-linguistic input. While MSG alleviates many of drawbacks of the purely text-based distributional models of meaning, it inherits some of their unrealistic learning assumptions. Specifically, in both MSG and text-based DSMs, the learner is asked to learn the meanings of word after processing corpora containing billions of words. Even simulations focusing on incremental learning tested performance changes in large steps (e.g., blocks of 100,000 words of running text in [9]). Humans, however, learn words one-by-one in an incremental fashion, and they can learn the meaning of new words from very limited exposure. Often, a single encounter suffices [105]. Can the distributional mechanisms shown by DSM simulations to be so effective in the long run also play a role in the initial stages of word learning from limited linguistic context?

There are *a priori* reasons to be doubtful that this would be the case. By its very

nature, distributional learning proceeds by progressive accumulation of association statistics. Single co-occurrences are not expected to be very telling, and it is only after extended periods of linguistic exposure that robust distributional patterns are expected to emerge. This is rather problematic if one wants to claim that distributional learning is a plausible mechanism employed by humans to learn word meaning, as it suggests that most day-to-day word learning must happen by different mechanisms, and only at a much later stage (and only for sufficiently frequent words) whatever has already been learned about a word is complemented with evidence from long-run distributional learning. Distributional learning thus starts looking like it is largely redundant with respect to other, more fundamental mechanisms. Indeed, *one-shot learning*, our ability to learn from just one example of a new word or concept, is often mentioned as a strong objection against the cognitive plausibility of distributional and more generally associationist learning mechanisms (e.g., [55], [105]). We think that this conclusion might be premature. Since discourses are generally coherent, even single word co-occurrences might often be at least broadly informative about a word meaning (a factor we will quantify below), making small-sample distributional learning generally reliable. Moreover, as a learner is progressively exposed to more language, we might observe a distributional bootstrapping effect, such that known words occurring as contexts of a new word help establishing a reasonable distributional profile for the new term quickly. Thus, in this study we decided to tackle, both experimentally and through a computational DSM simulation, the issue of whether it is possible for a learner to induce a reasonable semantic representation of a word from just very few (minimally 2, maximally 6) occurrences of the word in sentence contexts sampled to be representative of what one would encounter by chance in natural text (the minimum is set to 2 occurrences instead of 1 for technical reasons related to how we build our stimuli). Our results suggest that subjects (and DSMs) can indeed come up with at least a basic representation of a word meaning from this minimal amount of purely distributional evidence. We take this to be an important result that, combined with the many studies that have shown how distributional models of meaning naturally account for many linguistic and psychological phenomena, both at the ag-



gregate and at the item level, brings strong support for the view that distributional learning, specifically as implemented in DSMs, constitutes a plausible (although by no means unique) mechanism explaining our remarkable word learning capabilities.

Despite the theoretical importance of the issue, there is relatively little work focusing on fast learning of new word meanings from limited contexts. [65] proved that the distributional characteristics of short made-up linguistic contexts surrounding rare or nonce words significantly affect subjects' semantic similarity intuitions about these words. Recent ERP work has moreover shown that subjects can learn novel word meanings from minimal exposure to artificial sentences containing the words. Borovsky and colleagues [13, 12] found that a single exposure to a novel word within an informative sentence context (i) suffices to trigger plausibility effects (signaled by N400 patterns) concerning the felicity of the new word as a grammatical object of an existing verb (for example, after being presented with "*He tried to put the pieces of the broken plate back together with marf*", subjects displayed a semantic violation effect for "*She drove the marf*"); and (ii) the novel word primes a semantically related target (as measured by reduction in N400 amplitude) just like a known word would. Using similar methods, [67] showed that, when a nonce word is presented in highly informative sentence contexts, three exposures suffice for the word to prime semantically related terms in a way that is indistinguishable from the effect of a comparable real word. In all these experiments, contextual effects were triggered by embedding the novel word in carefully constructed sentences. Thus, the results constitute proof of principle that, in the presence of highly informative surrounding words, subjects can learn new word meanings from minimal exposure, but they do not establish the result beyond artificial contexts. Put differently, they do not tell us whether real linguistic contexts, of the sort that one would encounter during everyday reading or conversation, are informative enough for subjects to be able to exploit them for similarly rapid word learning.

There are two further potential shortcomings that often affect word learning studies. First, learning a new word is not the same as learning a new *meaning*. It is easy to think about situations in which we acquire new words for familiar meanings

(synonym learning). We may find out, for example, that the *buffalo* is also called *bison* (actually, the latter is a more proper term). One can also attach new meanings to familiar words, as in the case of polysemy. We may have originally known, for example, that *spam* denotes a type of canned meat, and only subsequently extended the word to denote junk e-mail. However, arguably, most often a speaker must learn how *new* words refer to *new* meanings (e.g., learning the word *iPad* the first time you heard about it, or saw one). That is, new word and concept acquisition typically proceed in parallel. Still, in much of the previous literature on word learning, including the studies we reviewed, participants are *de facto* asked to associate new labels to familiar meanings, that is, they are faced with a synonym learning task. In the example above, subjects must discover that *marf* is a new synonym of *glue* ([67], report that, in 91% of their informative-condition sentences, subjects were able to guess the real word masked by the nonsense string, or a closely related term).

Second, the relevant literature follows two main strategies to verify successful learning. A classic approach is to test if a subject can associate a new word with the right visual referent: e.g., the subject hears “Oh look, a *marf*” while seeing images of glue and a number of distractors, and the experiment measures whether she directs her gaze at glue [105]. Being able to pick the right referent is a fundamental aspect of learning the meaning of a (concrete) word, but, as [12] observe, “our lexical knowledge is often far richer than simple associations between labels, physical objects and features. Word representations are complex and multi-faceted. [...] The] meaning [of a novel word ...] must be appropriately situated within the local context and dynamic semantic landscape of the mental lexicon.” The studies discussed above address this criticism by probing purely language-based semantic properties of the novel word (e.g., whether it primes a related term). However, this method, in turn, gives up the link to visual referents as an important component of meaning. Especially when investigating the role of purely linguistic distributional evidence, it is fundamental instead to test whether subjects are not only getting lexical relations right (which might be explained away by relatively superficial textual co-occurrence effects), but also inferring the likely perceptual properties of the concept they are learning about.

In the present study, we adopt a novel experimental paradigm that is aimed at addressing these unrealistic assumptions of current DSM models. First, we do not employ artificially constructed texts to probe word learning. We extract natural sentences from existing corpora, substituting a target word with a non-word. Second, inspired by the perception literature, in which “chimeric” images are generated by mixing visual features of different objects or animals [84], we try to trigger genuine novel-meaning learning (as opposed to synonym matching) by mixing sentences referring to two related but distinct words. For example, the context of one of our lexical chimeras is a small random sample (that we call a *passage*) of sentences originally containing either the word *bear* or the word *gorilla*, both replaced by the non-word *mohalk*. Because gorillas and bears are related, the passage will paint a coherent picture of mohalks as large mammals, but subjects won’t be able to play a synonym guessing game, since there is no single word that *mohalk* is masking. Third, participants are asked to evaluate a series of probe items for their similarity to the novel term meaning, thus going beyond the simple labeling tasks criticized by Borovsky and colleagues. However, since we aim at testing both strictly linguistic and more general conceptual knowledge that subjects associate to the novel word, we employ as probe items both words (Experiment 1), and images (Experiment 2) of objects that are expected to be more or less related to the chimeric concept. Finally, after measuring human performance on our distributional learning task, we simulate it with a MSG, our novel DSM architecture combining linguistic and visual knowledge.

Our results confirm that (i) (adult) subjects can extract basic knowledge about the meaning of a new word, including both linguistic and visual aspects of the denoted concept, from a very small number of sentences randomly extracted from natural texts, and that (ii) a DSM exposed to the same kind of limited evidence induces multimodal knowledge about the new word in a way that is remarkably similar to what we observe in subjects.

The rest of the chapter is organized as follows: Section 3.2 describes our experimental materials and introduces the computational model. The two experiments are reported in sections 3.3 and 3.4, respectively. The final discussion in Section 5.4

summarizes our main results, and looks at our computational learning method from the broader perspective of word learning models.

## 3.2 Experimental setup

### 3.2.1 Experimental materials

We investigated the generation of novel concepts from short passages in adult human participants (see, e.g., [34], for general motivation to study word learning in adults). In particular, we created novel concepts, that we call *chimeras*, by combining two related but distinct words, the chimera *components*. A *passage* then consisted of a mixture of natural corpus sentences (minimally 2, maximally 6) that contained the two components, except that all instances of either component were masked by the same nonce word. Subjects were induced to believe that the sentences all pertained to the same novel concept, and they were tested for their intuitions about the meaning of this novel concept by being asked relatedness judgments between the latter and other terms, that we call *probes*. For example, Figure 3-2 below shows materials pertaining to the chimera with *gorilla* and *bear* as components. In one trial, illustrated in the figure, the shared nonce word masking the components was *mohalk*. Subjects (and the computational model) were presented a passage composed of 4 sentences where instances of *gorilla* and *bear* were replaced by *mohalk*, and immediately asked to rate the degree of relatedness of *mohalk* with, e.g., *dishwasher* or (in a separate trial) *lion*. As the figure shows, the probes were presented verbally in Experiment 1 and visually in Experiment 2.

#### Chimeras

To obtain the chimeras, we started with 33 basic-level concrete concepts that were also used by [30] (three of their concepts were excluded due to technical reasons). We call each of these concepts a *pivot*, and we match it with another concept, that we call a *compatible* term, by using the following procedure. For each pivot, we ordered

all the terms in the norms of [66] by similarity to the pivot, using the pre-compiled similarity table available with these concept norms (which is based on conceptual feature overlap, where concept features were elicited from subjects). We traversed the resulting ranked lists, picking as compatible term the first word that was not a synonym, close co-hyponym or hyper-/hyponym of the pivot. Whenever possible, we tried to match pivot and compatible terms with similar usage (e.g., both mass terms, or both likely to frequently occur in the plural) and reasonably similar in visual shape and function. No pivot was picked as compatible term of another pivot, and no compatible term was repeated across pivots. Except for *plane/ship*, the compatible term was always among the 10 most similar items to the pivot (for *plane*, the most similar concepts were birds). A chimera is given by the combination of a pivot and the corresponding compatible term (the chimera components). The first two columns of Table 3.1 report the full list of matched components forming our chimeras.

## Probes

Each chimera was associated with 6 words, whose degree of relatedness to the chimeras had to be assessed by the subjects based on the limited contexts surrounding the latter. We refer to such words as probe terms. We obtained probes spread across the relatedness spectrum by sampling them from different bins based on averaged McRae-norms similarity to chimera components. Examples of probes sampled in this way include *turnip* for the *corn/yam* chimera (in the top similarity bin according to the McRae norms) and *chipmunk* for *toaster/microwave* (bottom similarity bin).

In order to measure the degree of success of subjects and computational model in positioning a chimera in the right zone of conceptual space, we needed ground-truth chimera-probe relatedness (CPR) judgments. Since the chimeras are novel concepts, we estimated this quantity indirectly, by averaging similarity ratings between the (explicitly presented) chimera components and the probes, produced by a control group in a preliminary crowdsourcing experiment [85], using the Crowdfunder platform.<sup>1</sup> The ratings for the two components of a chimera with each probe were collected sep-

---

<sup>1</sup><http://www.crowdfunder.com>

arately. Subjects in the control group were given no cue that we intended to combine such ratings. Subjects were asked to rate the semantic relatedness of each chimera component word to the corresponding probes, for a total of 396 pairs (66 components by 6 probes), using a 7-point scale ranging from 1 (“completely unrelated”) to 7 (“almost the same meaning”). Ten judgments were collected for each pair. In the instructions, we stressed that we were interested in intuitions concerning word *meanings*, and that the relation between the element pairs could be more or less strict. The CPR of a chimera-probe pair was computed as the average rating of the two chimera components with the probe. For example, given the *bear/gorilla* chimera and the *lion* probe, we used averaged *bear-lion* and *gorilla-lion* ratings obtained from the control group as our CPR estimate.

Since participants’ semantic intuitions about chimeras in the experiments below were elicited through short text passages that pertain, in equal amounts, to the two components, it is reasonable to assume that the full chimera concept is an average of the component meanings (the *bear/gorilla* chimera denotes an animal somewhere in-between bears and gorillas), and that the “real” CPR score should be close to that obtained by averaging component-probe ratings. Note that one chimera’s components are very similar concepts. Consequently, they will in general have very comparable similarities to the probes (correlations between ‘component’-specific CPRs across all chimeras and probes are quite high:  $r = .73$  for word-to-word judgments and  $r = .69$  for the word-to-image judgments we use in Experiment 2 as discussed next). It is thus unlikely that the CPR scores of a chimera as a genuine separate concept would be very different from the means of the two component CPR scores. If subjects find both *bears* and *gorillas* very different from a *dishwasher*, it is probable that they would assign a similarly low *dishwasher* score to the *bear/gorilla* chimera, if such creature existed. Relatedness scores *averaged* across the two chimera components worked better as CPR estimates (in the sense of fitting our experimental results more closely) than relying on maximum or minimum component-to-probe relatedness. Even using the ground-truth scores for the most informative component in each passage (in the sense discussed below) did not lead to an improvement in model fit. To conclude, although

averaging component CPR scores can only provide an indirect estimate of a chimera’s CPR score, we have good reasons to believe that such estimate is quite accurate.

For Experiment 2, we replaced the probe words with images of the corresponding concepts. As seen in Figure 3-2 below, images were not artificially edited to represent the probe concepts in stereotypical settings. They were instead natural pictures of the relevant objects (that is, they were shot by amateur or professional photographers for their own purposes), taken from the ImageNet database we will briefly describe in Section 3.2.2.

Table 3.1 presents, for each chimera, its components together with the 6 probes and the respective control-group-generated word- and image-based CPR values. For example, the first row reports that one chimera was formed by the *alligator* pivot and by the compatible term *rattlesnake*, that the McRae similarity between these concepts is 0.39 (on a 0-1 scale), that *rattlesnake* is the second most similar term to *alligator* in the McRae norms (thus, it has rank 2 in the sorted neighbour list). It further shows that the *crocodile* probe received an average CPR of 4.15 (on a 1-7 scale) with this chimera when *crocodile* was presented as a word, and of 4.20 when it was presented as a picture, and similarly for all the other probes associated to the chimera.

## Passages

Each chimera was associated to 10 passages consisting of 6 sentences each (3 sentences per component). The latter were randomly extracted, in comparable proportions, from two corpora representative of written and spoken English (British National Corpus)<sup>2</sup> and Web texts (ukWaC, see Section 3.2.2 below), respectively (we avoided other commonly used textual sources likely to provide overly informative text, such as the Wikipedia). The sentences were manually checked to make sure they did not predicate mutually contradictory properties of the chimera (as the chimera components are related but different concepts). We also substituted sentences in which a component word was not used as intended (e.g., *bear* used as a verb). The sentences were not edited in any way. On average, they contained 17.6 words each. All occurrences of

---

<sup>2</sup><http://www.natcorp.ox.ac.uk>

Pivot	Compat.	Sim(Rank)	Probes							
			Word-based CPR				Image-based CPR			
alligator	rattlesnake	0.39 (2)	crocodile	iguana	gorilla	banner	buzzard	shovel		
			4.15 4.20	2.95 2.65	1.75 1.55	1.15 1.00	2.10 1.40	1.05 1.05		
bomb	missile	0.64 (1)	bazooka	gun	bayonet	bullet	buckle	canoe		
			3.60 3.25	3.10 2.55	2.40 2.50	3.85 3.10	1.20 1.10	1.20 1.00		
broccoli	spinach	0.74 (1)	celery	radish	grape	salamander	budgie	pot		
			2.80 2.70	2.85 2.30	2.45 1.80	1.40 1.00	1.45 1.00	1.15 1.70		
cannon	rifle	0.43 (2)	pistol	bomb	harpoon	trolley	lion	bagpipe		
			3.45 2.80	3.35 2.55	2.55 3.55	1.00 1.20	1.20 1.15	1.20 1.20		
car	van	0.60 (1)	skateboard	jeep	train	fridge	shed	parakeet		
			1.70 1.20	5.10 6.05	3.20 2.35	1.25 1.05	1.30 1.05	1.30 1.00		
caterpillar	cockroach	0.33 (1)	beetle	grasshopper	spider	shack	porcupine	crane		
			2.90 2.45	3.30 2.85	3.30 2.20	1.20 1.20	1.80 1.80	1.55 1.45		
cello	bagpipe	0.45 (8)	harmonica	drum	racquet	cabin	house	bolt		
			2.75 1.80	2.55 1.75	1.20 1.10	1.25 1.15	1.15 1.00	1.30 1.00		
clarinet	trombone	0.67 (2)	banjo	gorilla	whistle	worm	dress	pine		
			2.85 2.60	1.15 1.00	2.85 1.60	1.25 1.20	1.20 1.10	1.25 1.00		
corkscrew	grater	0.38 (5)	machete	hook	crane	rocket	bookcase	pickle		
			1.55 1.55	1.90 1.25	1.10 1.15	1.10 1.25	1.35 1.15	1.40 1.10		
corn	yam	0.29 (6)	turnip	eggplant	parsley	peach	buffalo	buggy		
			3.20 2.50	3.10 2.25	2.80 1.60	2.60 1.80	1.20 1.35	1.15 1.05		
cucumber	celery	0.66 (1)	rhubarb	onion	pear	strawberry	limousine	cushion		
			3.15 2.40	3.05 2.45	2.45 1.85	2.60 2.45	1.20 1.00	1.25 1.05		
dishwasher	oven	0.37 (4)	stove	microwave	kettle	cage	wastebin	leopard		
			3.60 4.70	3.85 3.40	2.10 1.75	1.15 1.00	1.45 1.10	1.20 1.00		
drums	tuba	0.55 (2)	bagpipe	harmonica	whistle	shotgun	bear	bouquet		
			3.05 2.60	2.70 1.90	2.65 1.60	1.20 1.00	1.05 1.10	1.10 1.00		
elephant	bison	0.46 (5)	caribou	groundhog	hare	spider	catapult	bolt		
			3.00 2.40	1.90 1.80	2.15 1.45	2.05 1.25	1.10 1.10	1.15 1.05		
gorilla	bear	0.42 (5)	lion	horse	chipmunk	bayonet	raisin	dishwasher		
			2.90 1.90	2.15 1.95	2.40 1.80	1.30 1.10	1.25 1.00	1.05 1.00		
guitar	harpsichord	0.76 (3)	tuba	racquet	barrel	shack	ladle	pumpkin		
			2.65 2.35	1.30 1.10	1.25 1.05	1.25 1.00	1.20 1.00	1.10 1.05		
harp	banjo	0.75 (3)	harpsichord	harmonica	peg	sled	shed	bedroom		
			3.80 2.95	3.20 1.95	1.25 1.00	1.40 1.05	1.35 1.05	1.10 1.00		
kettle	pan	0.23 (8)	toaster	bucket	hatchet	razor	pistol	helmet		
			2.25 1.70	2.45 1.30	1.25 1.10	1.15 1.05	1.05 1.05	1.20 1.10		
ladle	colander	0.51 (3)	tongs	corkscrew	kettle	stool	pencil	sellotape		
			2.70 1.85	1.55 1.40	2.50 1.40	1.10 1.05	1.25 1.00	1.10 1.10		
mittens	socks	0.51 (3)	sweater	boot	trousers	carpet	bra	corkscrew		
			2.15 1.50	2.35 1.80	2.25 1.85	1.25 1.80	2.20 1.35	1.20 1.00		
owl	partridge	0.61 (2)	duck	swan	jet	platypus	pig	armour		
			2.60 2.85	3.00 2.00	1.20 1.10	1.85 1.40	2.05 1.80	1.20 1.05		
peacock	goose	0.70 (2)	eagle	airplane	bear	ox	crane	cucumber		
			3.30 1.90	1.40 1.10	2.10 1.60	2.35 1.45	2.25 1.00	1.45 1.00		
piano	accordion	0.61 (3)	harpsichord	trumpet	typewriter	penguin	olive	lettuce		
			3.70 4.10	3.20 2.10	1.50 1.45	1.20 1.10	1.05 1.05	1.20 1.10		
plane	ship	0.20 (36)	jet	yacht	hawk	stork	corkscrew	nightgown		
			4.90 4.45	3.65 3.15	1.25 1.10	1.30 1.30	1.25 1.05	1.10 1.15		
potato	turnip	0.54 (1)	broccoli	onion	tomato	trout	cantaloupe	cork		
			2.90 1.95	3.05 2.30	2.85 2.25	1.70 1.30	2.30 3.00	1.20 1.00		
refrigerator	closet	0.38 (4)	cupboard	basement	mixer	dishwasher	ladle	boat		
			3.15 2.60	1.65 1.40	1.85 1.20	1.85 2.35	1.40 1.00	1.30 1.05		
saxophone	harmonica	0.49 (8)	clarinet	harp	buckle	wrench	urn	mackerel		
			3.60 2.75	3.25 2.25	1.20 1.00	1.05 1.00	1.45 1.10	1.15 1.05		
scarf	sweater	0.47 (2)	glove	robe	tie	swimsuit	nylons	spear		
			3.25 2.30	2.95 2.40	3.35 2.15	2.05 1.05	2.55 1.40	1.30 1.00		
scooter	skateboard	0.50 (5)	cart	jeep	boat	toy	pencil	tomato		
			2.35 1.30	2.15 1.15	1.65 1.40	3.65 1.00	1.00 1.00	1.05 1.00		
toaster	microwave	0.55 (3)	pot	apron	kettle	tongs	cheetah	chipmunk		
			1.90 1.85	1.50 1.50	2.35 2.05	1.85 1.30	1.15 1.05	1.20 1.05		
train	bus	0.35 (3)	jet	taxi	buggy	submarine	crane	grasshopper		
			2.70 2.05	2.70 2.45	2.35 2.45	2.15 1.35	1.40 1.60	1.10 1.00		
trouser	shirt	0.35 (2)	pants	shawl	cape	curtain	pajama	cart		
			4.25 4.15	2.90 1.60	2.50 2.55	1.50 1.20	3.35 3.90	1.15 1.15		
violin	flute	0.50 (6)	harp	drum	racquet	colander	rocket	radish		
			3.20 2.55	2.75 2.05	1.25 1.00	1.15 1.05	1.20 1.15	1.20 1.05		

Table 3.1: Chimeras and probes. For each chimera, the table reports, in this order, the chimera pivot and compatible components, their McRae similarity (and rank of compatible item among neighbours of pivot) and the probes with (left) word-based and (right) image-based CPR scores.



It was illustrated by a picture of Bernie Grant, the black Labour candidate,  
with the hairy body of a **MOHALK**. [*gorilla*]  
ANCHORAGE-- During September an unusually high number of **MOHALKS** were sighted  
in Alaska's North Slope oil fields. [*bear*]  
But watch out for nerds running up the down escalators, down the up escalators and generally  
acting as **MOHALKS** on crack. [*gorilla*]  
Of interest, during these cleaning activities he unearthed what appeared to be a **MOHALK**  
trap wedged in a crack. [*bear*]

Experiment 1

**DISHWASHER**

**LION**

Experiment 2



Figure 3-1: Example of the materials used in an experimental trial for the *gorilla/bear* chimera, corresponding to the nonce string MOHALK (the original word in each sentence, shown in squared brackets, was not presented to subjects). Unrelated and related probes are shown bottom left and right, respectively (in Experiment 1, only word probes were presented, in Experiment 2, only images).

either original component in a passage were replaced with the same non-word (for example, all the occurrences of *gorilla* and *bear* in a given passage were replaced by *mohalk*). The 330 non-words (33 chimeras by 10 passages) were generated using WUGGY [48]. They were 2- or 3-syllable-long nonsense strings respecting English orthotactics, and not containing productive affixes (e.g., *-ing*).

In order to quantify the amount of context that is needed to learn new meanings, we manipulated the number of sentences by creating shorter passages from the original ones. Three possible *passage length* levels were considered (2, 4, 6 sentences), each of them including an even number of sentences for either component (incidentally, this is the reason why we can't go down to 1 occurrence per chimera).

We calculated an *a-posteriori* estimate of passage informativeness with respect to the chimeras they contain. A passage is informative if it contains words that are descriptive of the chimera component concepts: e.g., a *bear/gorilla* passage containing words such as *fur* or *animal*, that are likely to be used in definitions or prototypical descriptions of bears and gorillas, is more informative than a passage that does not contain such words. Concretely, the informativeness score of a passage is the number

of distinct words in the passage that are identified as properties of the component concepts in the conceptual norms of [66]. We checked that informativeness was distributed not only across different passages, but also within each chimera. The average informativeness range was 0-4.5, going from *car-van* (with an informativeness range of 0-2) to *caterpillar-cockroach* and *guitar-harpsicord* (with an informativeness range of 0-7). Moreover, the correlation between passage length and informativeness was relatively low ( $r = .38$ ), indicating that one variable could not be reduced to the other.

One possible concern with our design is that many passages could be so informative about the chimera components that subjects could have guessed these items and based their ratings on them, rather than trying to infer a genuinely new meaning from the passage. To address this concern, we ran an additional crowdsourcing experiment via Crowdfunder. In this control experiment, we asked subjects if they could guess the word or words that were masked in each of the 330 6-sentence passages. A total of 101 subjects took part in the experiment. Each passage was presented to 10 subjects, and a single subject could maximally rate 50 passages. Note that, whereas the subjects of our main experiments, to be introduced below, were instructed to think of the passages as involving a new concept, the current control group was explicitly instructed to try to guess existing words for the masked slots. Moreover, unlike the main experiment subjects, the control group was told that different sentences in the same passages could involve different masked words, and they were allowed to list more than one word per passage. Given these differences, this control experiment is testing a worst-case scenario, in which no main-experiment subject followed our instructions, they decided to use a word-guessing strategy instead, and they also figured out that the passages might refer to different concepts. Despite the facilitated setup, for only 17% of the passages more than half of subjects were able to guess at least one of the two chimera components. However, in many of these cases the subjects guessed a chimera component simply because they produced a list of related concepts, that included the component. For example, for an *harmonica/saxophone* passage, a subject wrote *harp, guitar, bass, saxophone*, suggesting he/she understood that the passage referred

to a musical instrument, but he/she was not really able to tell which one. But this is exactly the sort of effect we are after, with subjects being able to tell that the “new” *harmonica/saxophone* chimera is an instrument, without being able to map it precisely to an existing one. The proportion of passages in which more than half the subjects were able to guess at least one chimera component without also producing irrelevant concepts is 10%. There is no single passage for which more than half the subjects were able to guess both chimeras.

### 3.2.2 Computational model

#### Model training

Simulating the acquisition of adult-like competence, we pre-train MSG-B (see Chapter 2) on ukWaC, a text corpus of about 2 billion running words from random Web pages.<sup>3</sup> We associate all corpus occurrences of 3,825 distinct concrete words (about 5% of the total word tokens) with their visual representations. These words are selected as follows: They must have an entry in ImageNet (see next paragraph), occur at least 500 times in ukWaC and have concreteness score  $\geq 0.5$  according to [106]. There were 5,100 words matching these constraints, but 25% of them (1,275) were set apart to estimate the mapping function described in the next section. In total, 116 out of the 155 probes and 50 out of the 66 chimera components were associated to visual representations during training.

Visual representations are obtained by randomly sampling 100 images labeled with the word of interest from ImageNet [23], a large database of “natural” pictures downloaded from the Web and annotated with words. We automatically extract a 4096-dimensional vector from each image using the Caffe toolkit [43], together with the state-of-the-art convolutional neural network of [54]. Convolutional neural networks, taking inspiration from biological vision, extract multiple layers of increasingly abstract features from images. Our vectors correspond to activation on the top (fc7) layer of the network, that captures complex, gestalt-like shapes [121]. For compu-

---

<sup>3</sup><http://wacky.sslmit.unibo.it>

tational reasons, we reduce the vector to 300 dimensions through Singular Value Decomposition. The visual representation of a word is simply the centroid of the vectors of the 100 images labeled with the word.

The following hyperparameters are fixed without tuning: word vector dimensionality: 300; subsampling:  $s=0.001$ ; window size:  $c=5$ . The following hyperparameters are tuned on the text9 corpus:<sup>4</sup>  $k=5$ ,  $\gamma=0.5$ ,  $\lambda=0.0001$ . The remaining model parameters are estimated by back-propagation of error via stochastic gradient descent on the input corpus and associated visual representations.

### Simulating relatedness judgments

Given the passage in which a chimera occurs, simulating the chimera-to-probe (CPR) relatedness judgments requires the generation of a vector representation for the chimera (that, recall, from the participants’ point of view, is a novel word only occurring in the passage). Taking inspiration from early work on contextualizing word meaning in DSMs [86], we represent a chimera in a sentence context as the centroid of the MSG vectors of all other (known) words in the sentence. The centroids obtained from each sentence in a passage are normalized and averaged to derive the passage-based chimera vector. This approach can leverage more information (harnessing the full vectors of all words in the passage) than simply updating a new chimera vector in a single step of the MSG prediction task.

In Experiment 1, we quantify the degree of relatedness of chimera and probe word by measuring the cosine formed by the angle of the chimera vector (constructed as just described) and the pre-trained MSG vector of the probe word.

In Experiment 2, we need to measure the relatedness of the chimera with an *image* probe. To represent the probe, we extract a visual feature vector from the image presented to subjects with the same convolutional neural network used for MSG training (see previous section). Since the probe and the chimera vectors lay in different spaces (visual and linguistic, respectively), their mutual relatedness cannot be directly assessed as in Experiment 1. The two representations could be connected

---

<sup>4</sup><http://mattmahoney.net/dc/textdata.html>

in two ways: by estimating a linguistic equivalent of the probe visual vector by means of a vision-to-language mapping function, or, alternatively, inducing a visual vector from the chimera linguistic vector through the inverse mapping. We report results obtained with the former procedure because the resulting relatedness scores were more in line with human judgments.

Specifically, we learn to map visual representations onto linguistic space. The mapping, consisting of weight matrix  $\mathbf{M}_{v2w}$ , is induced with a max-margin ranking loss objective, analogously to what is done in MSG estimation. The loss for pairs of visual and linguistic training items  $(\mathbf{x}_i, \mathbf{y}_i)$  and the corresponding mapping-derived predictions  $\hat{\mathbf{y}}_i = \mathbf{M}_{v2w}\mathbf{x}_i$  is defined as

$$\sum_{j \neq i}^k \max\{0, \gamma - \cos(\hat{\mathbf{y}}_i, \mathbf{y}_i) + \cos(\hat{\mathbf{y}}_i, \mathbf{y}_j)\} \quad (3.1)$$

where  $\gamma$  and  $k$  are tunable hyperparameters denoting the margin and the number of negative examples, respectively. The optimal hyperparameters  $\gamma = 0.7$  and  $k = 2$  were chosen on a set of 300 validation concepts from the training set. To train the mappings, we use word and visual representations derived from 1,275 concrete concepts (selected as described in the previous section). We estimate the mapping parameters  $\mathbf{M}_{v2w}$  with stochastic gradient descent and per-parameter learning rates tuned with AdaGrad [24].

### 3.3 Experiment 1: Estimating relatedness of chimeric concepts to word probes

#### 3.3.1 Methods

##### Participants

Participants were recruited through crowdsourcing, and in particular using the Crowdfunder platform. As is standard for Crowdfunder studies, participants were free to abandon a job at any time, and there were no restrictions on how many jobs a par-

It was illustrated by a picture of Bernie Grant, the black Labour candidate,  
with the hairy body of a **MOHALK**. [*gorilla*]  
ANCHORAGE-- During September an unusually high number of **MOHALKS** were sighted  
in Alaska's North Slope oil fields. [*bear*]  
But watch out for nerds running up the down escalators, down the up escalators and generally  
acting as **MOHALKS** on crack. [*gorilla*]  
Of interest, during these cleaning activities he unearthed what appeared to be a **MOHALK**  
trap wedged in a crack. [*bear*]

Experiment 1

**DISHWASHER**

**LION**

Experiment 2



Figure 3-2: Example of the materials used in an experimental trial for the *gorilla/bear* chimera, corresponding to the nonce string MOHALK (the original word in each sentence, shown in squared brackets, was not presented to subjects). Unrelated and related probes are shown bottom left and right, respectively (in Experiment 1, only word probes were presented, in Experiment 2, only images).

ticipant could perform. The subjects eventually participating in the experiment were 213. As an average, each chimera was rated by 143 distinct subjects.

## Procedure

Relatedness ratings between the chimeras in each passage and the associated probes were collected. Passage length and probes were counterbalanced across items using a Latin-square design, so that a given combination of passage and probe appeared only once in each list. Each of the resulting 18 lists (6 probes by 3 passage lengths) included 330 items (33 chimeras by 10 passages). Each list was administered as a separate Crowdfunder job, resulting in a full within-subject design (that is, participants were presented all possible levels of factorial predictors, and a representative distribution of continuous predictors). In order to avoid large familiarity effects between different lists, jobs were presented in separate days. Moreover, non-words were randomized across lists, so that each non-word was associated to a different passage in each list.

Participants were asked to judge how related a familiar word (the probe) was to the concept denoted by the unknown word in the presented sentence. An example

trial is presented in Figure 3-2 (ignore the images). Participants were told that the unknown word could be thought of as referring to a new or very rare concept, or to a thing that might exist in the parallel universe of a science fiction saga. We also specified that the relation could be stricter or looser, and that we were specifically interested in the word *meanings*, rather than forms or sounds. Relatedness was rated on a 5-point scale. Seven judgments were collected for each item.<sup>5</sup>

## Data analysis

Human ratings were analyzed through mixed-effects models [4] including random intercepts for subjects and chimeras. The main predictor was CPR. In the analysis we also included covariates as potential modulators of this effect. We considered the interaction of CPR with passage informativeness. Moreover, passage length was also made to interact with CPR, to assess whether the raw amount of contextual information contributes to the quality of the generated novel representation. All predictors were mean-centered. We started from a full factorial model, and removed effects that did not contribute significantly to the overall fit. Having identified the best model, atypical outliers were identified and removed (employing 2.5 SD of the residual errors as criterion). The models were then refitted to ensure that the results were not driven by a few overly influential outliers. Statistics of the refitted models are reported.

To assess the performance of MSG, we ran the same statistical analysis using as dependent variable MSG-generated passage-based chimera-probe similarity scores in place of human intuitions. The only difference with respect to the analysis of human behavior is that, in the case of model predictions, there are no subjects as aggregation unit, resulting in a simpler random effect structure including only the chimera random intercept. The rationale behind this approach is that a good computational model should not (only) be correlated with human scores, but, more importantly, reproduce the statistical structure subtending behavioral measures: for example, if we observe a

---

<sup>5</sup>The full datasets of Experiment 1 and Experiment 2 are available from: <http://clic.cimec.unitn.it/Files/PublicData/chimeras.zip>

particular interaction in the analysis of participants’ response, this very same pattern should be observed in the computational simulation predictions (see [5], for a similar approach). Still, we also report direct correlation between average human ratings and model predictions.

As an additional control, we further tested the data by including random slopes in the statistical models. These random parameters are meant to capture the variability of a given fixed effect (e.g., CPR) between a series of aggregator units (e.g., subjects), and permit to evaluate whether the considered fixed effect is reliably observed across the considered units. Random slopes of each included predictor were tested in association with chimeras for the analysis on model data, and in association with chimeras and subjects for the analysis on human data. Random slopes were included in the model only if their contribution to the overall model fit was significant. This was tested by a goodness-of-fit test comparing the model before and after the inclusion of the additional random parameter.

### 3.3.2 Results

Table 3.2 presents a summary of the data collection results for the first experiment. Figure 3-3 represents the association between the two dependent variables (human ratings and model predictions) with respect to the observed data points. This direct correlation is at a highly significant  $r = .39$  ( $p = .0001$ ). Figure 3-4 represents the association between CPR and human responses (left-hand panel) and model predictions (right-hand panel).

Table 3.3 and Figure 3-5 report the results of the mixed-effects analysis on human responses and model predictions. As a result of the outlier-removal procedure, 3.4% of the data points were excluded from the human-rating dataset, and 2.3% of the data points were excluded from the model-prediction dataset.

Subject results are reported in the first three columns of Table 3.3. We observe a significant interaction between CPR and informativeness: the more informative the presented sentences are, the more markedly subject intuitions match ground-truth CPR. This is also evident from the left panel of Figure 3-5, representing the



	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5	Probe 6
Low inf.	2.65	2.56	2.39	2.31	2.06	2.02
Medium inf.	2.75	2.73	2.41	2.29	2.07	2.07
High inf.	2.87	2.72	2.42	2.27	2.08	2.03
2 sentences	2.71	2.69	2.42	2.28	2.07	2.02
4 sentences	2.74	2.68	2.41	2.28	2.06	2.05
6 sentences	2.83	2.67	2.40	2.31	2.09	2.05

Table 3.2: Rating distributions in Experiment 2. Average ratings (in different columns) are reported for set of probes grouped and ranked by their ground-truth CPR, crossed with passage informativeness (upper table) and passage length (lower table).

Predictor	Human responses			Model predictions		
	b	SEM	t	b	SEM	t
Intercept	2.138	0.066	32.44	0.166	0.007	21.87
Informativeness	0.019	0.005	5.73	0.008	0.001	10.49
CPR	0.305	0.003	65.96	0.056	0.001	51.27
Informativeness * CPR	0.031	0.004	8.48	0.008	0.001	9.56

Table 3.3: Experiment 1 (word probes): Results when analyzing either human responses or model predictions.

interaction captured in the mixed-effects analysis. Different lines in the plot represent the effect of (mean-centered) CPR at different level of informativeness: the higher the informativeness, the steeper the slope, the more aligned human responses are to ground-truth CPR. Moreover, the plot shows how the effect of CPR is evident also for trials in which informativeness is very low. This crucially indicates that participants are able to learn something meaningful from contexts that are not only short, but also poor in explicit evidence. Passage length (that is, the number of sentences forming the context) has no significant modulation on participants’ performance (and consequently it is not part of the final model): even when only two sentences are presented, the CPR effect does not significantly decrease.

The next set of columns in Table 3.3 reports the results when testing the MSG scores for passage-based chimera representations and probes. The pattern is close to what observed for human responses: we found no reliable effect of passage length, and a significant interaction between CPR and informativeness. The latter is also

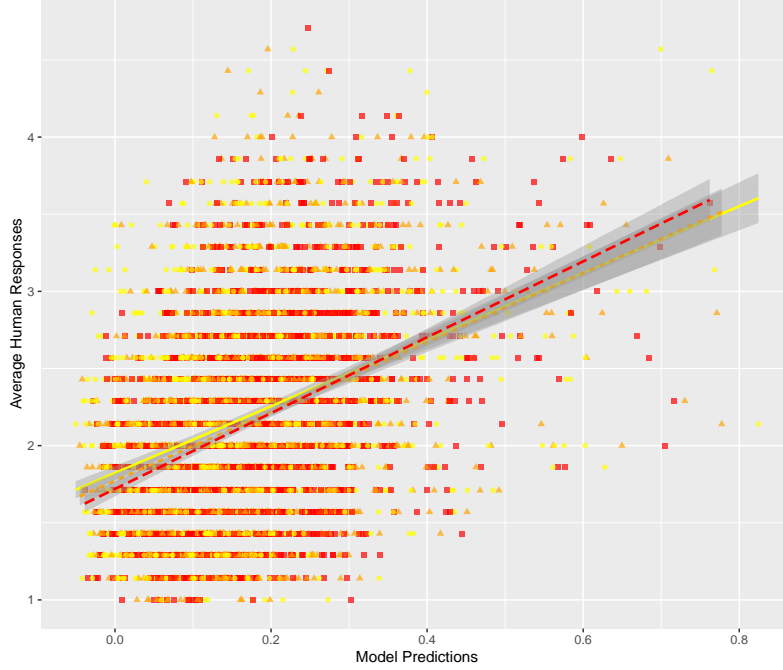


Figure 3-3: Experiment 1 (word probes): association between by-item average human responses and model predictions. Yellow circles: 2-sentence passages; orange triangles: 4-sentence passage; red squares: 6-sentence passages.

qualitatively very similar to the one found for human participants (compare left and right panel of Figure 3-5).

The mixed-effects models significantly improved following the inclusion of CPR random slopes for subjects and chimeras (human rating analysis) and for chimeras (model prediction analysis). However, the reported effects (Table 3.3) still hold following the inclusion of random slopes, and the pattern of results remains consistent with the one reported in Figure 3-5.

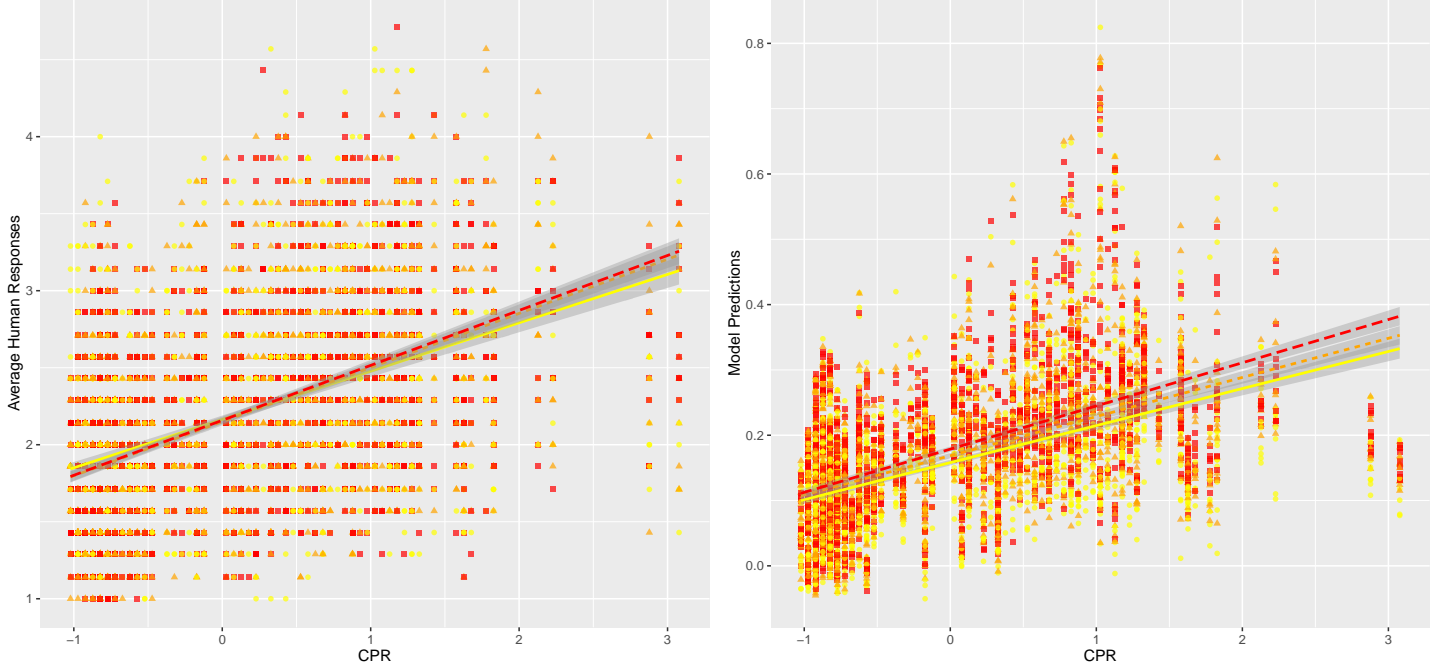


Figure 3-4: Experiment 1 (word probes): association between CPR and by-item average human responses (left-hand panel) and between CPR and model predictions (right-hand panel). Yellow circles: 2-sentence passages; orange triangles: 4-sentence passage; red squares: 6-sentence passages.

## 3.4 Experiment 2: Estimating relatedness of chimeric concepts to image probes

### 3.4.1 Methods

#### Participants

As for Experiment 1, participants were recruited through the Crowdfunder platform. Eventually, 168 subjects participated in Experiment 2. As an average, each chimera was rated by 120 distinct subjects.

#### Procedure

We used the same materials and followed the same procedure as in Experiment 1 above, running 18 separate Crowdfunder jobs on as many lists generated through a

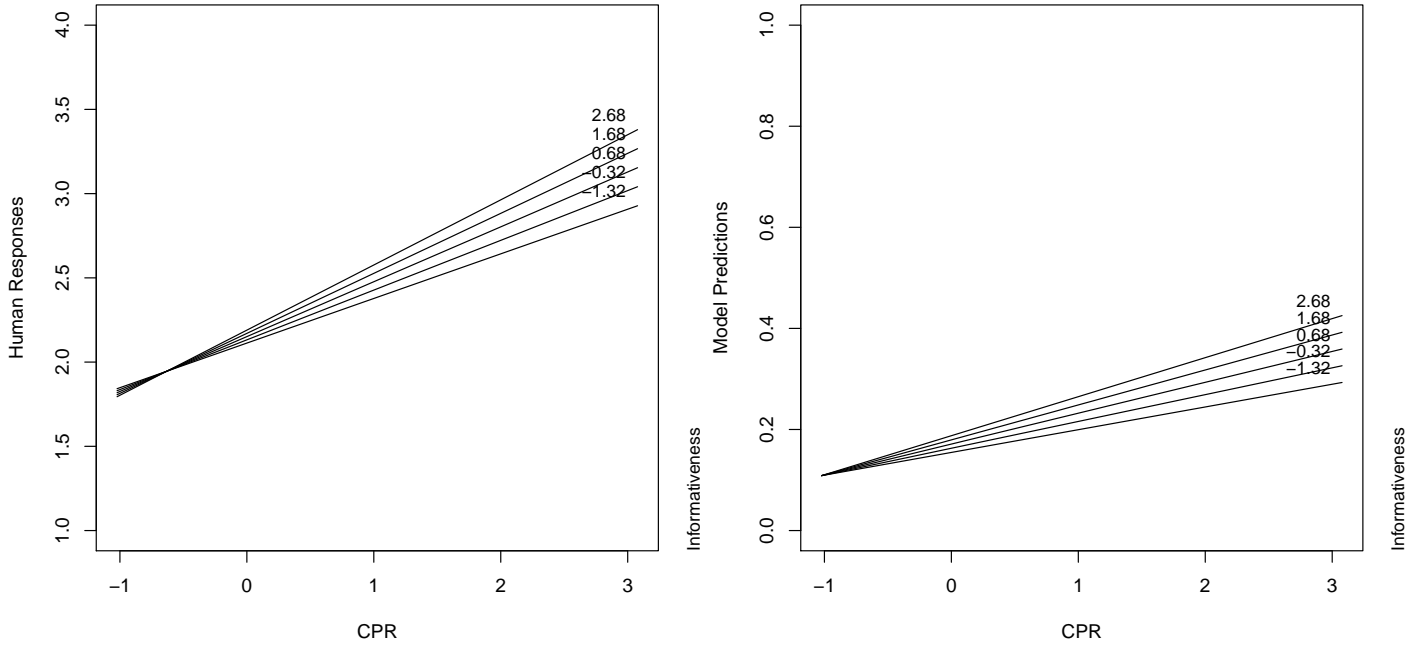


Figure 3-5: Experiment 1 (word probes): Interaction between informativeness and CPR for human responses (left panel) and model predictions (right panel).

Latin-square design. The only difference is that probe words were substituted with images of the corresponding concepts, and ground-truth CPR scores (see Section 3.2.1 above) based on component-word-to-probe-image similarity judgments.

Participants were asked to rate the relatedness between the meaning denoted by the unknown word and the presented image. In the instructions, we stressed the visual aspect of the task, asking participants to base their ratings on their impression, from the passage, of how the unknown concept would look like.

## Data Analysis

The data were analyzed as described above for Experiment 1.

	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5	Probe 6
Low inf.	2.52	2.33	2.23	2.12	1.87	1.85
Medium inf.	2.59	2.37	2.19	2.05	1.86	1.82
High inf.	2.71	2.49	2.19	2.01	1.82	1.82
2 sentences	2.59	2.39	2.20	2.08	1.86	1.86
4 sentences	2.60	2.41	2.20	2.02	1.85	1.82
6 sentences	2.65	2.40	2.20	2.06	1.84	1.79

Table 3.4: Rating distributions in Experiment 2. Average ratings (in different columns) are reported for set of probes grouped and ranked by their ground-truth CPR, crossed with passage informativeness (upper table) and passage length (lower table).

### 3.4.2 Results

Table 3.4 presents a summary of the results of the data collection for the second experiment. Figure 3-6 represents the association between the two dependent variables with respect to the observed data points. The correlation between them is at  $r = .34$  ( $p = .0001$ ). Figure 3-7 represents the association between CPR and human responses (left-hand panel) and model predictions (right-hand panel).

Table 3.5 and Figure 3-8 report the results of the mixed-effects analysis on human responses and model predictions. As a result of the outlier-removal procedure, 1.6% of the data points were excluded from the human-rating dataset, and 2.1% of the data points were excluded from the model-prediction dataset.

Results parallel Experiment 1. First, we observe significant effects of CPR, informativeness and their mutual interaction. Second, the modulation of passage length did not hold against the statistical controls, and was hence removed from the model. Third, the overall pattern observed in human judgments and model predictions is similar.

Also in this case, the mixed-effects models significantly improved following the inclusion of random slopes. Both CPR and informativeness random slopes (associated to both subjects and chimeras in the human rating analysis, and to chimeras in the model prediction analysis) significantly improved the model fit. Again, this more complex random structure does not change the overall pattern of the reported effects.

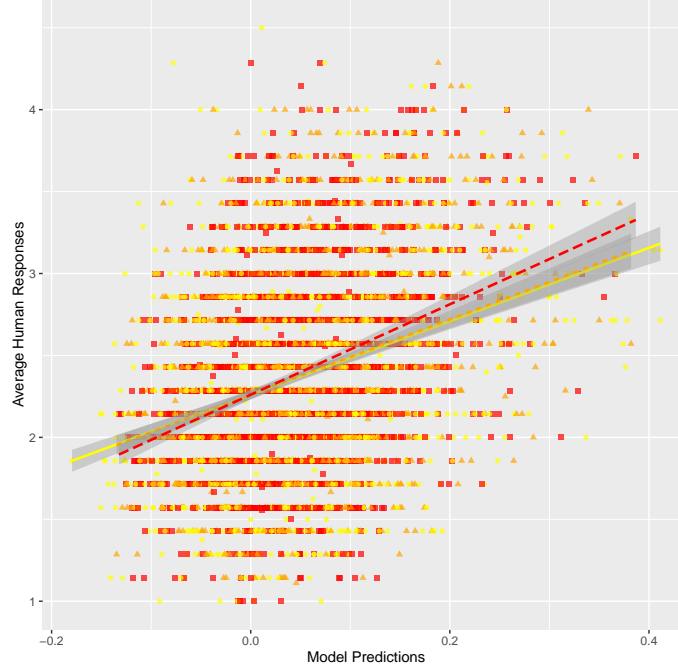


Figure 3-6: Experiment 2 (image probes): association between by-item average human responses and model predictions. Yellow circles: 2-sentence passages; orange triangles: 4-sentence passage; red squares: 6-sentence passages.

One advantage of a model closely mimicking subject behaviour, such as MSG, is that it can offer insights on the inner workings of word meaning formation. As an example of the possibilities offered by computer simulations, Figure 3-9 visualizes how MSG “imagines” the thing denoted by a novel word, based solely on the word representation it extracted from a single passage. We stress that what we report here is just an informal analysis of these visualizations, providing preliminary qualitative insight into how distributional concept formation might work, and we leave a systematic study using this methodology to future work.

To generate the images in Figure 3-9, we first obtained a visual vector by mapping the relevant passage-based chimera representation to visual space (using a mapping function from linguistic to visual representations analogous to the inverse one described in Section 3.2.2 above). Then, we retrieved the 50 pictures in our image pool (Section 3.2.2) whose visual feature vectors were nearest to the mapped chimera, and we superimposed them. Intuitively, if we have never seen an object, we might

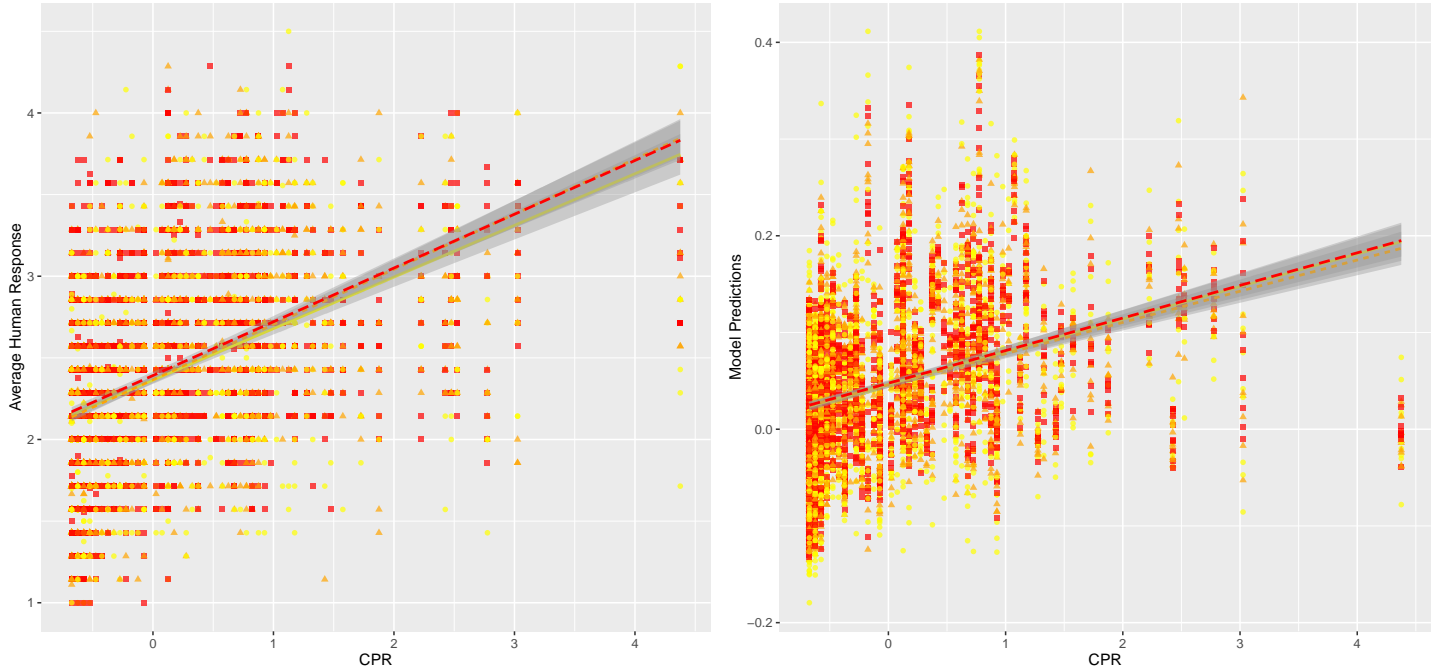


Figure 3-7: Experiment 2 (image probes): association between CPR and by-item average human responses (left-hand panel) and between CPR and model predictions (right-hand panel). Yellow circles: 2-sentence passages; orange triangles: 4-sentence passage; red squares: 6-sentence passages.

associate it to a mental image that is an average of those objects that we think are similar: Before ever seeing a *papaya*, we might expect it to look somewhere in between a *mango* and a *banana* (see [60], for further details on this image generation method).

The first two images in the figure correspond to two-sentence passages about the *caterpillar/cockroach* chimera. The first passage was “easy” for subjects (in the sense that averaged passage-based judgments across probes were highly correlated with ground-truth CPR,  $r = 0.95$ ), and indeed, when deriving its representation of the novel word from this context, MSG visualizes a brownish blob in the middle of green, suggesting a very sketchy bug. The second passage is harder (subjects-CPR correlation:  $r = 0.46$ ), and MSG is misled by the metaphorical usage in the second sentence, visualizing a city landscape. The third passage (pertaining to the *potato/turnip* chimera) is relatively easy for subjects ( $r = 0.71$ ), as well as for MSG, that is clearly pushed by the first sentence to emphasize the edible aspect of the

Predictor	Human responses			Model predictions		
	b	SEM	t	b	SEM	t
Intercept	2.245	0.056	39.98	0.045	0.006	7.81
Informativeness	0.014	0.004	3.75	0.004	0.001	6.08
CPR	0.346	0.006	56.93	0.033	0.001	31.05
Informativeness * CPR	0.039	0.005	8.55	0.006	0.001	7.45

Table 3.5: Experiment 2 (image probes): Results when analyzing either human responses or model predictions.

chimera. The fourth passage was particularly difficult ( $r = 0.07$ ), and again we see how MSG is also misled (the image seemingly dominated by *peppermint* and/or *field*).

### 3.5 Discussion

Despite the fact that distributional learning is typically seen as a long-term process based on large-corpus statistics, our experimental and computational results suggest that this mechanism supports the construction of a reasonable semantic representation as soon as a new word is encountered.

Our experimental data confirmed that purely linguistic distributional evidence constitutes a precious source of information in acquiring word meaning, and that very limited amounts of uncontrolled text, of the sort one naturally encounters in everyday life, suffice for human participants to form reasonably accurate semantic representations, that involve both intuitions about the position of the concepts denoted by the new words in language-based semantic space (Experiment 1), and predictions about their visual aspect (Experiment 2).

Contrary to expectations about evidence accumulation in incremental learning, the length of the passage presented to subjects had no effect on the quality of the representations they extracted. The length difference of our stimuli was limited, ranging from 2 to 6 sentences, and it is likely that more marked differences will significantly impact human performance. Still, the present results indicate that, when evidence is very limited, human learning is mostly determined by the quality of the supplied information (captured here by the informativeness measure), rather than by



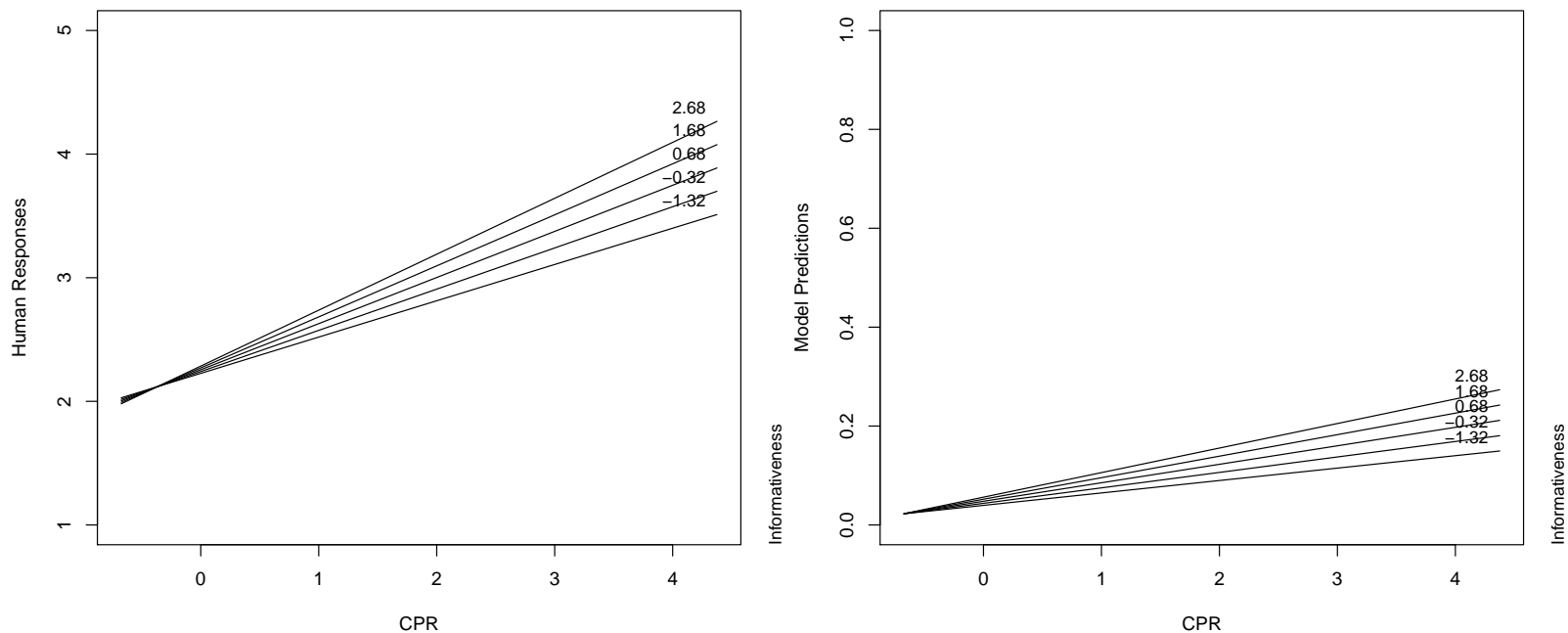


Figure 3-8: Experiment 2 (image probes): Interaction between informativeness and CPR on human responses (left panel) and model predictions (right panel).

its quantity.

A version of the MSG distributional semantic model, that we equipped with adult-like semantic competence by pre-training it on natural linguistic and image data, performed the word learning tasks in ways that are remarkably close to those of human subjects. Indeed, the reported simulations mirror the interaction between informativeness and CPR observed in our participants, suggesting that the model is not just generating cognitively plausible representations, but also exploiting, in this process, the same type of information that guides human learning.

Our results suggest that the simple context-prediction mechanisms encoded in MSG might suffice to explain the human ability to extract conceptual representations from minimal textual contexts. Specifically, MSG has been trained to produce word representations that are highly predictive of the linguistic contexts in which words occur, and of the visual features of the objects that they denote. Such architecture,

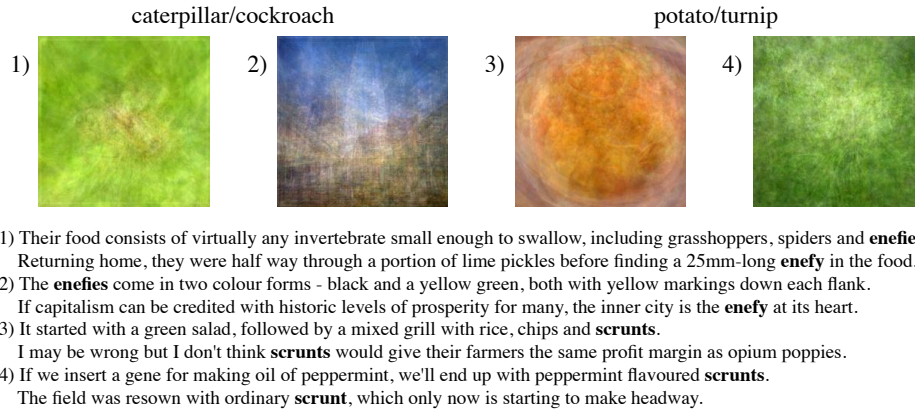


Figure 3-9: How MSG visualizes the novel word in each of 4 passages, two constructed from *caterpillar/cockroach* contexts, and two from *potato/turnip* contexts.

when presented with a new word in a passage, can immediately derive a linguistic representation for the word by interpolating over the known words in the context. This produces a representation for the new term that can be used to estimate its degree of relatedness to other words. Moreover, through cross-modal mapping, the model can simulate human intuitions about how the new concept must look like. In this perspective, the MSG architecture plausibly illustrates how previously acquired knowledge (both linguistically and visually connoted) is exploited to generate novel representations when a new word is first encountered.

The feature-level associationist learning process implemented in MSG can be seen as complementary to the widely studied cross-situational statistics tracking methods needed to link words to their referents [95]. The MSG learning procedure assumes that the right referent of (a subset of) words is known, and it relies on visual properties of referents (as well as linguistic contexts) to build word meaning representations. It is thus natural to assume that MSG learning operates on input that has been referentially disambiguated through standard cross-situational word-referent linking.

On the other hand, MSG captures aspects of learning that are problematic for classic cross-situational models. For example, it can explain why *similarity* and *context* play an important role in the computation and retention of cross-situational statistics [109]: similar instances of an object will have similar visual features, speeding up learning. It can explain why word learning, as [95] put it, is “incremental and

slow”, and simply associating a new word to a new referent in an unambiguous context (“fast mapping”) is not guarantee of genuine word learning, that is, full understanding and long-term retention of the word meaning. In the MSG model, the meaning (= vector) of a word is updated at each encounter with the word based on linguistic and visual contexts, as well as indirect influence from other words whose meaning is being constructed in the same representational space. Thus, meaning is indeed acquired incrementally: As our experiments showed, the model can already build a rough sketch of a word meaning from very few exposures, but the representation will keep becoming more precise (leading to better linguistic and cross-modal prediction) as more evidence is seen. Finally, by learning also from purely linguistic contexts, MSG accounts for abstract words, or even learning about concrete terms in absence of a referent, which remains a mystery for standard cross-situational learning models.

In the future, we intend to pursue the issue of whether word-level cross-situational learning and feature-level MSG-style learning have to be treated as connected but distinct processes, or if they can instead be unified into a single computational architecture.



# Chapter 4

## Attentive Social Multimodal Skip-gram Model

### 4.1 Introduction

In the previous chapters we introduced MSG, a model of semantic learning that enriches linguistic vectors with extra-linguistic information captured in real images. While we showed that MSG is an effective learner, even in case of minimal exposure to natural text, it makes strong assumptions not found in real learning scenarios. First, so far we assumed that we know the right object to be associated with a word, i.e., when somebody utters “cat” our learner knows exactly where to look in order to extract extra-linguistic information for learning the meaning of “cat”. However, this is clearly unrealistic during language acquisition, since children apriori do not know the meaning of words (moreover, cats do not carry labels!). Second, MSG induced word meaning by being presented with arbitrarily chosen sentences and objects that did not belong to a real learning episode, i.e., (Wikipedia) sentences were paired with (ImageNet) objects even though the former were not uttered in the presence of the latter.

In this chapter, we address all these limitations and present the ATTENTIVE SOCIAL MULTIMODAL SKIP-GRAM MODEL. Like the original MSG, our model learns multimodal word embeddings by reading an utterance sequentially and making, for each

word, two sets of predictions: (a) the preceding and following words, and (b) the visual representations of objects co-occurring with the utterance. However, unlike MSG, we do not assume we know the right object to be associated with a word. We consider instead a more realistic scenario where multiple words in an utterance co-occur with multiple objects in the corresponding scene. Under this referential uncertainty, the model needs to induce word-object associations as part of learning, relying on current knowledge about word-object affinities as well as on any social clues present in the scene. Moreover, the objects to be presented alongside text are not chosen arbitrarily. Instead, we use CHILDES Database [63], a corpus of transcribed interactions between children and their caregivers, that contains (among others) multi-modal information in the form of which objects were present during these interactions (see Section 4.4).

Finally, while multi-modal semantic learning has focused so far on incorporating extra-linguistic information extracted from images of objects, in this chapter we take a step towards incorporating *social cues* that are found in real communicative episodes between people, i.e., eye-gaze, gestures, body posture. Social cues reflect speakers' intentions and generally contribute to the unfolding of the communicative situation [100]. Moreover, in the context of language learning it, children who engage in joint attention (e.g. looking at particular objects together) with their caregivers have shown learn words faster [18]. As a first step towards developing full-fledged learning systems that leverage all signals available within a communicative setup, in our extended model we incorporate information regarding the objects that caregivers are holding.

## 4.2 Related Work

Computational models of word learning typically approximate the perceptual context that learners are exposed to through artificial proxies, e.g., representing a visual scene via a collection of symbols such as `cat` and `dog`, signaling the presence of a cat, a dog, etc. [119, 27, *inter alia*].<sup>1</sup> While large amounts of data can be generated in this way,

---

<sup>1</sup>See [44] for a recent review of this line of work, and another learning model using, like ours, real visual input.

they will not display the complexity and richness of the signal found in the natural environment a child is exposed to.

There is however work on learning from multimodal data [82, 118, a.o.] as well as work on learning distributed representations from child-directed speech [9, 51, a.o.], to the best of our knowledge ours is the first method which learns distributed representations from multimodal child-directed data. For example, in comparison to [118]’s model, our approach (1) induces distributed representations for words, based on linguistic and visual context, and (2) operates entirely on distributed representations through similarity measures without positing a categorical level on which to learn word-symbol/category-symbol associations. This leads to rich multimodal conceptual representations of words in terms of distributed multimodal features, while in Yu’s approach words are simply distributions over categories. It is therefore not clear how Yu’s approach could capture phenomena such as predicting appearance from a verbal description or representing abstract words—all tasks that our model is at least in principle well-suited for. Note also that [29]’s Bayesian model we compare against could be extended to include realistic visual data in a similar vein to Yu’s, but it would then have the same limitations.

Our work is also related to research on reference resolution in dialogue systems, such as [47]. However, unlike Kennington and Schlangen, who explicitly train an object recognizer associated with each word of interest, with at least 65 labeled positive training examples per word, our model does not have any comparable form of supervision and our data exhibits much lower frequencies of object and word (co-)occurrence. Moreover, reference resolution is only an aspect of what we do: Besides being able to associate a word with a visual extension, our model is simultaneously learning word representations that allow us to deal with a variety of other tasks—for example, as mentioned above, guessing the appearance of the object denoted by a new word from a purely verbal description, grouping concepts into categories by their similarity, or having both abstract and concrete words represented in the same space.

### 4.3 Attentive Social MSG Model

Similar to the standard skipgram, the model’s parameters are context word embeddings  $\mathbf{W}'$  and target word embeddings  $\mathbf{W}$ . The model aims at optimizing these parameters with respect to the following multi-task loss function for an utterance  $w$  with associated set of objects  $U$ :

$$L(w, U) = \sum_{t=1}^T (\ell_{ling}(w, t) + \ell_{vis}(w_t, U)) \quad (4.1)$$

where  $t$  ranges over the positions in the utterance  $w$ , such that  $w_t$  is  $t^{\text{th}}$  word. The linguistic loss function is the standard skip-gram loss [68]. The visual loss is defined as:

$$\ell_{vis}(w_t, U) = \sum_{s=1}^S \lambda \alpha(\mathbf{w}_t, \mathbf{u}_s) g(\mathbf{w}_t, \mathbf{u}_s) + (1 - \lambda) h(\mathbf{u}_s) g(\mathbf{w}_t, \mathbf{u}_s) \quad (4.2)$$

where  $\mathbf{w}_t$  stands for the column of  $\mathbf{W}$  corresponding to word  $w_t$ ,  $\mathbf{u}_s$  is the vector associated with object  $U_s$ , and  $g$  the penalty function

$$g(\mathbf{w}_t, \mathbf{u}_s) = \sum_{\mathbf{u}'} \max(0, \gamma - \cos(\mathbf{w}_t, \mathbf{u}_s) + \cos(\mathbf{w}_t, \mathbf{u}')), \quad (4.3)$$

which is small when projections to the visual space  $\mathbf{w}_t$  of words from the utterance are similar to the vectors representing co-occurring objects, and at the same time they are dissimilar to vectors  $\mathbf{u}'$  representing randomly sampled objects. The first term in Eq. 4.2 is the penalty  $g$  weighted by the current word-object affinity  $\alpha$ , inspired by the “attention” of [6]. If  $\alpha$  is set to a constant 1, the model treats all words in an utterance as equally relevant for each object. Alternatively it can be used to encourage the model to place more weight on words which it already knows are likely to be related to a given object, by defining it as the (exponentiated) cosine similarity between word and object normalized over all words in the utterance:

$$\alpha(\mathbf{w}_t, \mathbf{u}_s) = \frac{\exp(\cos(\mathbf{w}_t, \mathbf{u}_s))}{\sum_r \exp(\cos(\mathbf{w}_r, \mathbf{u}_s))} \quad (4.4)$$



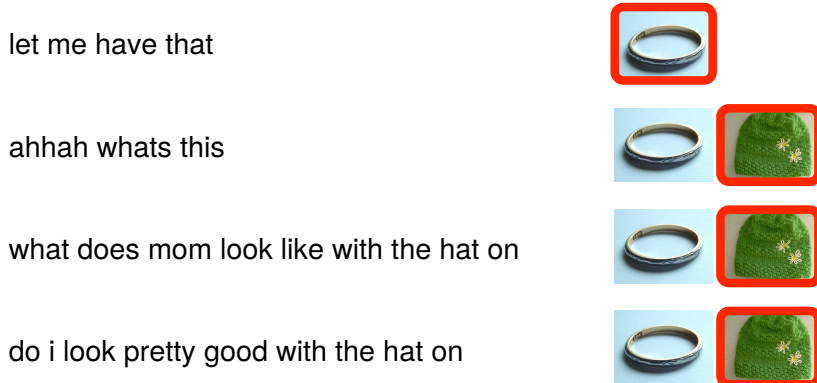


Figure 4-1: Fragment of the IFC corpus where symbolic labels `ring` and `hat` have been replaced by real images. Red frames mark objects being touched by the caregiver.

The second term of Eq. 4.2 is the penalty weighted by the social salience  $h$  of the object, which could be based on various cues in the scene. In our experiments we set it to 1 if the caregiver holds the object, 0 otherwise.

We experiment with three versions of the model. With  $\lambda = 1$  and  $\alpha$  frozen to 1, the model reduces to the original MSG, but now trained with referential uncertainty. The ATTENTIVE MSG sets  $\lambda = 1$  and calculates  $\alpha(\mathbf{w}_t, \mathbf{u}_s)$  using Equation 4.4 (we use the term “attentive” to emphasize the fact that, when processing a word, the model will pay more attention to the more relevant objects). Finally, ATTENTIVE SOCIAL MSG further sets  $\lambda = \frac{1}{2}$ , boosting the importance of socially salient objects.

All other hyperparameters are set to the values found by [59] to be optimal after tuning, except hidden layer size that we set to 200 instead of 300 due to the small corpus (see Section 4.4). We train the MSG models with stochastic gradient descent for one epoch.

## 4.4 The Illustrated Frank et al. Corpus

[29] present a Bayesian cross-situational learning model for simulating early word learning in first language acquisition. The model is tested on a portion of the Rollins section of the CHILDES Database [63] consisting of two transcribed video files (`me03` and `di06`), of approximately 10 minutes each, where a mother and a pre-verbal infant

play with a set of toys. By inspecting the video recordings, the authors manually annotated each utterance in the transcripts with a list of object labels (e.g., `ring`, `hat`, `cow`) corresponding to all midsize objects judged to be visible to the infant while the utterance took place, as well as various social cues. The dataset includes a gold-standard lexicon consisting of 36 words paired with 17 object labels (e.g., *hat*=`hat`, *pig*=`pig`, *piggie*=`pig`).<sup>2</sup>

Aiming at creating a more realistic version of the original dataset, akin to simulating a real visual scene, we replaced symbolic object labels with actual visual representations of objects. To construct such visual representations, we sample for each object 100 images from the respective ImageNet [23] entry, and from each image we extract a 4096-dimensional visual vector using the Caffe toolkit [43], together with the pre-trained convolutional neural network of [54].<sup>3</sup> These vectors are finally averaged to obtain a single visual representation of each object. Concerning social cues, since infants rarely follow the caregivers’ eye gaze but rather attend to objects held by them [120], we include in our corpus only information on whether the caregiver is holding any of the objects present in the scene. Note however that this signal, while informative, can also be ambiguous or even misleading with respect to the actual referents of a statement. Figure 4-1 exemplifies our version of the corpus, the Illustrated Frank et al. Corpus (IFC).

Several aspects make IFC a challenging dataset. Firstly, we are dealing with language produced in an interactive setting rather than written discourse. For example, compare the first sentence in the Wikipedia entry for *hat* (“A *hat* is a head covering”) to the third utterance in Figure 4-1, corresponding to the first occurrence of *hat* in our corpus. Secondly, there is a large amount of referential uncertainty, with up to 7 objects present per utterance (2 on average) and with only 33% of utterances explicitly including a word directly associated with a possible referent (i.e., not taking into account pronouns). For instance, the first, second and last utterances in Figure 4-1 do not explicitly mention any of the objects present in the scene. This uncertainty also

---

<sup>2</sup><http://langcog.stanford.edu/materials/nipsmaterials.html>

<sup>3</sup>To match the hidden layer size, we average every  $k = 4096/200$  original non-overlapping visual dimensions into a single dimension.

<i>Model</i>	<i>Best-F</i>
MSG	.64 (.04)
ATTENTIVMSG	.70 (.04)
ATTENTIVESOCIALMSG	.73 (.03)
ASMSG+shuffled visual vectors	.65 (.06)
ASMSG+randomized sentences	.59 (.03)
BEAGLE	.55
PMI	.53
BAYESIAN CSL	.54
BEAGLE+PMI	.83

Table 4.1: Best-F results for the MSG variations and alternative models on word-object matching. For all MSG models, we report Best-F mean and standard deviation over 100 iterations.

extends to social cues: only in 23% of utterances does the mother explicitly name an object that she is holding in her hands. Finally, models must induce word-object associations from minimal exposure to input rather than from large amounts of training data. Indeed, the IFC is extremely small by any standards: 624 utterances making up 2,533 words in total, with 8/37 test words occurring only once.

## 4.5 Experiments

We follow the evaluation protocol of [29] and [52]. Given 37 test words and the corresponding 17 objects (see Table 4.2), all found in the corpus, we rank the objects with respect to each word. A mean *Best-F* score is then derived by computing, for each word, the top F score across the precision-recall curve, and averaging it across the words. MSG rankings are obtained by directly ordering the visual representations of the objects by cosine similarity to the MSG word vectors.

Table 4.1 reports our results compared to those in earlier studies, all of which did not use actual visual representations of objects but rather arbitrary symbolic IDs. BAYESIAN CSL is the original Bayesian cross-situational model of [29], also including social cues (not limited, like us, to mother’s touch). BEAGLE is the best semantic-space result across a range of distributional models and word-object match-

ing methods from [52]. Their distributional models were trained in a *batch mode*, and by treating object IDs as words so that standard word-vector-based similarity methods could be used to rank objects with respect to words. PLAIN MSG is outperforming nearly all earlier approaches by a large margin. The only method bettering it is the BEAGLE+PMI combination of Kievit-Kylar et al. (PMI measures direct co-occurrence of test words and object IDs). The latter was obtained through a grid search of all possible model combinations performed directly on the test set, and relied on a weight parameter optimized on the corpus by assuming access to gold annotation. It is thus not comparable to the untuned MSG.

PLAIN MSG, then, performs remarkably well, even without any mechanism attempting to track word-object matching across scenes. Still, letting the model pay more attention to the objects currently most tightly associated to a word (ATTENTIVEMSG) brings a large improvement over PLAIN MSG, and a further improvement is brought about by giving more weight to objects touched by the mother (AttentiveSocialMSG). As concrete examples, PLAIN MSG associated the word *cow* with a pig, whereas ATTENTIVEMSG correctly shifts attention to the cow. In turn, ATTENTIVESOCIALMSG associates to the right object several words that ATTENTIVEMSG wrongly pairs with the hand holding them, instead.

One might fear the better performance of our models might be due to the skip-gram method being superior to the older distributional semantic approaches tested by [52], independently of the extra visual information we exploit. In other words, it could be that MSG has simply learned to treat, say, the *lamb* visual vector as an arbitrary signature, functioning as a semantically opaque ID for the relevant object, without exploiting the visual resemblance between lamb and sheep. In this case, we should obtain similar performance when arbitrarily shuffling the visual vectors across object types (e.g., consistently replacing each occurrence of the *lamb* visual vector with, say, the *hand* visual vector). The lower results obtained in this control condition (ASMSG+shuffled visual vector) confirm that our performance boost is largely due to exploitation of genuine visual information.

Since our approach is incremental (unlike the vast majority of traditional distri-

butional models that operate on batch mode), it can in principle exploit the fact that the linguistic and visual flows in the corpus are meaningfully ordered (discourse and visual environment will evolve in a coherent manner: a hat appears on the scene, it’s there for a while, in the meantime a few statements about hats are uttered, etc.). The dramatic quality drop in the ASMSG+randomized sentences condition, where ATTENTIVESOCIALMSG was trained on IFC after randomizing sentence order, confirms the coherent situation flow is crucial to our good performance.

**Minimal exposure.** Given the small size of the input corpus, good performance on the word-object association already counts as indirect evidence that MSG, like children, can learn from small amounts of data. In Table 4.2 we take a more specific look at this challenge by reporting ATTENTIVESOCIALMSG performance on the task of ranking object visual representations for test words that occurred *only once* in IFC, considering both the standard evaluation set and a much larger confusion set including visual vectors for 5.1K distinct objects (those of [59]). Remarkably, in all but one case, the model associates the test word to the right object from the small set, and to either the right object or another relevant visual concept (e.g., a ranch for *moocows*) when the extended set is considered. The exception is *kitty*, and even for this word the model ranks the correct object as second in the smaller set, and well above chance for the larger one. Our approach, just like humans [105], can often get a word meaning right based on a single exposure to it.

**Generalization.** Unlike the earlier models relying on arbitrary IDs, our model is learning to associate words to actual feature-based visual representations. Thus, once the model is trained on IFC, we can test its generalization capabilities to associate known words with new object instances that belong to the right category. We focus on 19 words in our test set corresponding to objects that were normed for visual similarity to other objects by [90]. Each test word was paired with 40 ImageNet pictures evenly divided between images of the gold object (*not* used in IFC), of a highly visually similar object, of a mildly visually similar object and of a dissimilar

<i>word</i>	<i>gold object</i>	<i>17 objects</i>		<i>5.1K objects</i>	
		<i>nearest</i>	<i>r</i>	<i>nearest</i>	<i>r</i>
bunny	bunny	bunny	<b>1</b>	bunny	<b>1</b>
cows	cow	cow	<b>1</b>	lea	<b>7</b>
duck	duck	duck	<b>1</b>	mallard	<b>4</b>
duckie	duck	duck	<b>1</b>	mallard	<b>3</b>
kitty	kitty	book	<b>2</b>	bookcase	<b>66</b>
lambie	lamb	lamb	<b>1</b>	lamb	<b>1</b>
moocows	cow	cow	<b>1</b>	ranch	<b>4</b>
rattle	rattle	rattle	<b>1</b>	rattle	<b>1</b>

Table 4.2: Test words occurring only once in IFC, together with corresponding gold objects, AttentiveSocialMSG top visual neighbours among the test items and in a larger 5.1K-objects set, and ranks of gold object in the two confusion sets.

one (for *duck*: *duck*, *chicken*, *finch* and *garage*, respectively). The pictures were represented by vectors obtained with the same method outlined in Section 4.4, and were ranked by similarity to a test word ATTENTIVESOCIALMSG representation.

Average Precision@10 for retrieving gold object instances is at 62% (chance: 25%). In the majority of cases the top-10 intruders are instances of the most visually related concepts (60% of intruders, vs. 33% expected by chance). For example, the model retrieves pictures of sheep for the word *lamb*, or bulls for *cow*. Intriguingly, this points to classic overextension errors that are commonly reported in child language acquisition [80].

## 4.6 Discussion

Our very encouraging results suggest that multimodal distributed models are well-suited to simulating human word learning. We think the most pressing issue to move ahead in this direction is to construct larger corpora recording the linguistic and visual environment in which children acquire language, in line with the efforts of the Human Speechome Project [83, 81]. Having access to such data will enable us to design agents that acquire semantic knowledge by leveraging all available cues present in multimodal communicative setups beyond visual context, such as learning agents

<i>word</i>	<i>gold</i>	<i>nearest</i>	<i>Accuracy top10</i>
bear	bear	bear	20%
book	book	book	70%
books	book	book	70%
cow	bull	cow	50%
cows	cow	cow	50%
moocow	bull	cow	50%
moocows	cow	cow	60%
bird	duck	duck	50%
birdie	duck	duck	70%
duck	duck	duck	70%
duckie	duck	duck	70%
lamb	lamb	lamb	70%
lambie	lamb	lamb	70%
mirror	mirror	mirror	20%
oink	pig	pig	90%
pig	pig	pig	80%
piggie	pig	pig	70%
piggies	pig	pig	70%

Table 4.3: Generalization experiment. Search space of each concept containing 40 elements. 10 gold, 10 visually similar, 10 less similar, 10 random

that can automatically predict eye-gaze [79] and incorporate this knowledge into the semantic learning process.





# Chapter 5

## Towards Multi-Agent Communication-Based Language Learning

### 5.1 Introduction

One of the most ambitious goals of AI is to develop intelligent conversational agents able to communicate with humans and assist them in their tasks. Thus, solving language processing is a key step toward such agents. In the previous chapters, we aimed at designing computational learners that learn language without strong supervision (see ATTENTIVE SOCIAL MSG presented in Chapter 4) within communicative episodes in the form of recorded multimodal interactions between children and their caregivers. However, training on *canned* conversations does not allow these “passive” learners to experience the interactive aspects of conversations. While this paradigm is a good way to learn general statistical associations (in the case of ATTENTIVE SOCIAL MSG, these would consist in associations between words and their referents in the external world) it focuses on the structure of language rather than on its purpose, that is, the fact that we use words to make things happen [3]!

In the language community, after the seminal work on the “blocks-world” environ-

ment of [115], we are now experiencing a revival of interest in language learning frameworks that are centered around communication and interaction (e.g., the *Roadmap* of [69] or the more recent *dialogue-based learning* proposal of [112]). Similar trends are also taking place in other fields of Artificial Intelligence, witness the revival of interest in *game playing* with the recent ground-breaking results of DQN on Atari [73] and AlphaGo [91], following precursors such as DeepBlue [17] and TD-Gammon [104].

Focusing on language, current approaches to communication-based language learning simulate interactive environments in diverse ways, e.g., by having agents interacting directly with humans or other scripted agents. Both approaches exhibit potentially important limitations. The *human-in-the-loop* approach (e.g., the SHRLDU program of Terry Winograd, robots learning via interacting with humans as in [103]) faces serious scalability issues, as active human intervention is obviously required at each step of training. Scripted Wizard-of-Oz environments [69] shift the burden of heavy manual engineering from the learning agent to designing the right behaviour for the programmed teaching agents.

In this chapter, we are proposing a radically different research program, namely *multi-agent communication-based language learning within a multimodal environment*. The essence of this proposal is to let computational agents co-exist, so that their co-existence constitutes the interactive environment. In this multi-agent environment, agents need to collaborate to perform a task, and we hypothesize that (with the right priors and constraints) developing language production and understanding will be prerequisites to successful communication. Note that we are not suggesting that the “passive” setup should be abandoned, as, even in the interactive paradigm, large-scale statistical learning is expected to be important, e.g., to let agents discover how to produce grammatical sentences, how to recognize object categories in general or even how to provide generic descriptions of what is present in a scene. Still, interaction is required for many other tasks, that only make sense within a communicative setup, e.g., how to refer to specific things, or how to ask a good question or respond to it.

Several points make this proposal attractive. First of all, this framework requires minimum human intervention for designing agents, the environment and its physics,

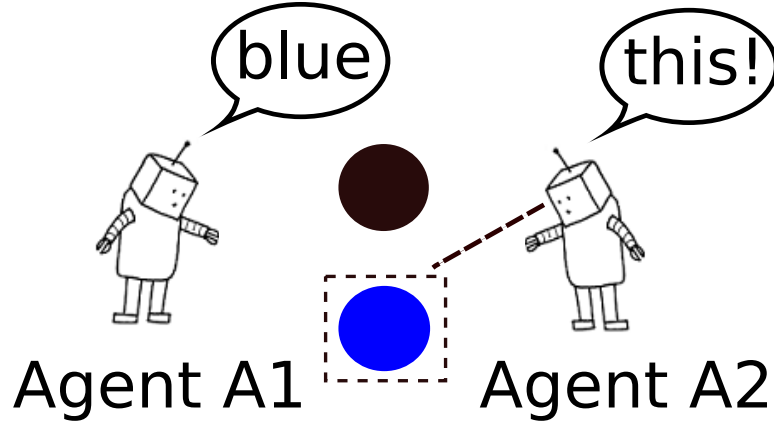


Figure 5-1: Schematic representation of the referential game. Agent A1 has to describe the dashed object with an attribute. Agent A2, has to guess which object A1 was referring to. In this example, the agents agree on the referent, so the communicative act was successful.

e.g, rewards, although humans do still need to specify the nature of the tasks that agents need to perform. Computational agents will co-exist (co-operate or antagonize) and self-organize freely, interacting with each other and being encouraged to learn in order to achieve communication. For example, imagine the simple case in which an agent needs to have some object that some other agent possesses, and she starts asking for it in various ways. Only when she manages to make herself understood she will be able to get hold of that object. The sort of learning taking place in such setup will have to be based on *active* request for information, and it will probably foster incremental agreement by *interaction*.

We start by considering the most basic act of communication, namely referring to things. We design multimodal riddles in the form of *referential games* (see Figure 5-1) [33]. The speaker in this game is asked to refer to one of the visible objects by uttering an expression. The listener, who sees the same objects but has no knowledge regarding which object the speaker was asked to describe, needs to identify it based on the speaker's expression. Similar simulations with signaling games have been studied in traditional literature of the emergence of language (see, e.g., [111] and [94] for surveys). While there is much there to be learned for our situation, given the focus of these works on studying the origins of language, the simulations from this period were rather restricted with respect to the nature and size of of input data,

i.e, mostly artificial input and small networks. In this work, we are arguing in favor of a multi-agent communication framework as the foundation of building machines that can communicate with us, thus focusing on more realistic data/input and less assumptions about their nature.

Importantly, the agents start in a *tabula rasa* state. They do not possess any form of language or understanding. They have no prior notion of the semantics of words. Meanings are assigned to words (that is, the arbitrary symbols used in our initial simulations) by playing the game and are reinforced by communication success. Thus, agents can agree to any sort of conceptualization and assign to any word any kind of interpretation that help them effectively solve the tasks. This essentially aligns with the view of [116] that *language meaning is derived from usage*.

We will report next a set of pilot experiments showing that, while it is feasible, within this multi-agent environment, to learn efficient communication protocols succeeding in the referential game, such protocols might not necessarily be aligned with the sort of semantics that exists in natural language. Interestingly, in the seminal language evolution experiment “Talking Heads” of [99], when two robots were left free to interact with each other, they developed an artificial language bearing little resemblance to natural language. Thus, we anticipate that, if we want to move forward with this research programme, grounding the agents’ communication into natural language will be crucial, since our ultimate goal is to be able to develop agents capable of communication with humans.

## 5.2 A two-agent referential game simulation

### 5.2.1 The game

We propose a simple referential game with 2 agents,  $A1$  and  $A2$ . The game is defined as follows:

- $A1$  is shown a visual scene with two objects and is told to describe the referent with a word that constitutes the referring expression ( $RE$ ).

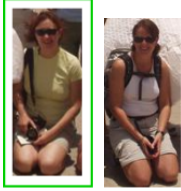

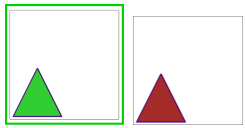
	ReferIt	Objects	Shapes
<b>visual scenes</b>			
<b>gold terms</b>	yellow	hamster	limegreen
<b>#unique images</b>	22.5k	46k	2.4k
<b>#visual scenes</b>	25k	495k	100k
<b>#gold terms</b>	3467	100	18

Table 5.1: Examples of  $\langle \text{referent}, \text{context} \rangle$  pairs from our 3 datasets. Referents are marked with a green square.

- A2 is shown the same visual scene without information on which is the referent, and given the RE has to “point” to the intended object
- if A2 points to the correct object, then both agents receive a game point.

Note that this game resembles the ReferItGame [46] played by humans and designed to collect RE annotations of real scenes.

We observe that this is a *co-operative* game, i.e., our agents must work together to achieve game points. A1 should learn how to provide accurate expressions that discriminate the object from all others, and A2 should be good at interpreting A1’s REs in the presence of the objects, in order to point to the correct one. Thus, with this game A1 and A2 learn to perform *referring expression generation* and *reference resolution* respectively.

### 5.2.2 Visual Scenes

Our current setup consists of visual scenes with only two objects, the *referent* and the *context*. Towards this end, we created 3 games from 3 different datasets by controlling the type of objects being paired in the visual scene. Each dataset focuses on some particular aspect of the referential game.

While our framework does not require obtaining gold annotations for the RE of the referent, we apply a number of heuristics to annotate each  $\langle \text{referent}, \text{context} \rangle$

pair with *gold words* acting as REs. This will allow us to conduct various analyses regarding the nature of the semantics that the agents assign to the induced words. Table 5.1 exemplifies an instance of the game for the different scenarios, and reports descriptive statistics for the 3 datasets.<sup>1</sup>

**ReferIt** The first game scenario uses data derived from the ReferItGame [46]. In its general format, ReferItGame contains annotations of bounding boxes in real images with referring expressions produced by humans when playing the game. In order to create plausible visual situations consisting of two objects only, we synthesize scenes by pairing each referent (as denoted by a bounding box in the image) with a distractor context that comes from the *same* image (i.e., some other bounding box in the image). Each bounding box in the initial ReferItGame is associated with a RE, which we pre-process to eliminate stop words, punctuation and spatial information, deriving single words attributes. We then follow a heuristic to obtain the *gold words* acting as the referring expression for a given  $\langle \text{referent}, \text{context} \rangle$  pair, by selecting words that were produced to describe the intended referent but not the context. The rationale for this decision is that a *necessary* condition for achieving successful reference is that REs accurately distinguish the intended referent from any other object in the context [21]. For maximizing the quality of the generated gold words, (i) we disregarded any distractor context whose bounding box overlapped significantly with the referent’s bounding box and (ii) we disregarded distractor contexts that had full word overlap with the referent, thus resulting in a null referring expression for the  $\langle \text{referent}, \text{context} \rangle$  pair.

**Objects** To control for the complexity of the visual scenes and words while maintaining real images, we created a simpler dataset in which referent and context are always different objects. For a list of 100 concrete objects ranging across different categories (e.g., *animal*, *furniture* etc), we synthesized  $\langle \text{referent}, \text{context} \rangle$  pairs by taking all possible combinations of objects and, for each object, sampling an image

---

<sup>1</sup>Our datasets will be made available.

from the respective ImageNet [23] object label entry. The RE/gold words for a given pair is then straightforwardly obtained by using the object name of the referent.

**Shapes** Finally, we introduce a third dataset which controls for the complexity that real images have, while allowing referent and context to differ in a diverse number of words, exactly like in real scenes. Following [1], we created a geometric shapes dataset consisting of images that contain a single object. We generate such single-object images by varying the values of 6 types of attributes and follow a similar approach as in the **ReferIt** dataset to annotate  $\langle$ referent, context $\rangle$  pairs with gold words (i.e., by taking the difference of the attributes in the referent and context).<sup>2</sup>

### 5.2.3 Agent Players

**Agent A1 (Referring Expression Generation)** Agent A1 is performing a task analogous to *referring expression generation*. Unlike traditional REG research [22, 72, 46] that produces phrases or words as RE, in our current framework, agent A1 learns to predict a single word that discriminates the referent from the context. Note however that word meaning is not pre-defined. Instead, it emerges “on-demand” via their usage in the referential game. In particular, ideally agent A1 will learn to associate the words to systematic configurations of lower-level perceptual features present in the images. In the current experiments, the words are simply represented by numerical indices, but it would of course be trivial to associate such indices with phonetic strings.

Figure 5-2 illustrates the network architecture of A1. The model is presented with the two images that constitute a visual scene. We assume that agents are already equipped with a pre-trained visual system that converts the raw pixel input of the referent and the context to higher-level visual vectors  $\mathbf{v}$  (i.e., activation patterns on a high layer of a pre-trained convolutional network. For games using the **ReferIt** and **Objects** dataset we used the pre-trained VGG-network [92]. For the **Shapes**, given

---

<sup>2</sup>The 18 attributes grouped by type: shape= $\langle$ triangle, square, circle $\rangle$ , border color= $\langle$ fuchsia, indigo, crimson, cyan, black, limegreen, brown, gray $\rangle$ , horizontal position= $\langle$ up, down $\rangle$ , vertical position= $\langle$ right, left $\rangle$ , shape size= $\langle$ small, medium, big $\rangle$

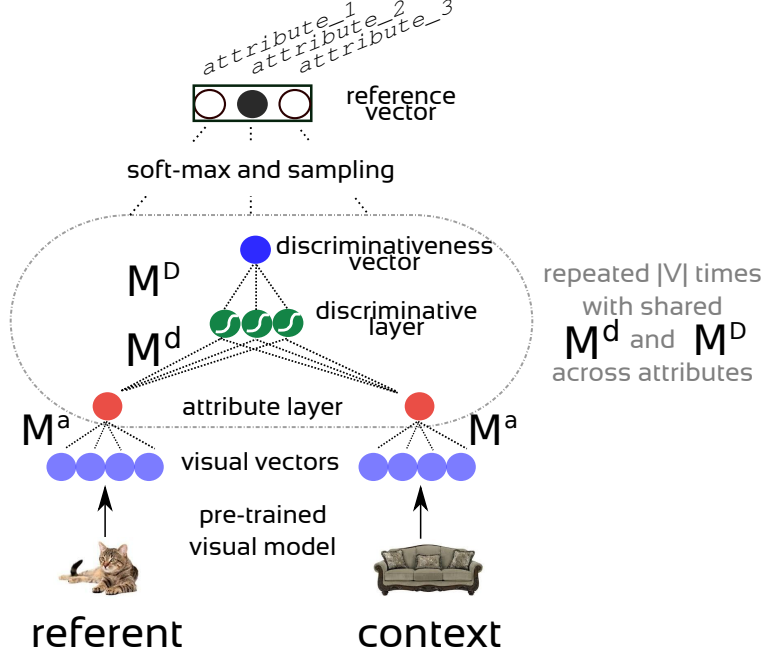


Figure 5-2: Neural network of player A1. In this particular game, A1 produces the reference vector that activates `word_2` as the RE, which is going to be passed over to A2.

that the nature of the images is different from that of usual ImageNet data used to train these networks, we trained our own model. Specifically, we trained a smaller network, i.e., AlexNet [54] on predicting bundles of two words. For both models, we use the second-to-last fully connected layer to represent images corresponding to 4096-D vectors. These visual vectors are mapped into *word* vectors of dimensionality  $|V|$  (cardinality of all available words, where this vocabulary size parameter is decided by the experimenter), with weights  $M^a \in \mathbf{R}^{4096 \times |V|}$  shared across the two objects. Intuitively, this layer learns which words are active for specific objects.

*Pairwise* interactions between word vectors of referent and context are captured in the *discriminative* layer. This layer, processes, for each word, the two units, one for each object, and by applying a linear transformation with weights  $M^d \in \mathbf{R}^{2 \times h}$ , followed by a sigmoid activation function, finally derives a single value by another linear transformation with weights  $M^D \in \mathbf{R}^{h \times 1}$ , producing  $d_v$  which encodes the degree of discriminativeness of word  $v$  for the specific referent. The same process with the same shared weights  $M^d$  and  $M^D$ , across words is applied to all words  $v \in V$ ,



to derive the estimated discriminativeness vector  $\mathbf{d}$ . Finally, the discriminativeness vector is converted to a probability distribution, from which the player samples one word  $a$  acting as the referring expression. This word is encoded in the *reference vector* with one-hot like representation, and is passed over to A2. The learnable parameters of A1 are  $\theta_{A1} = \langle \mathbf{M}^a, \mathbf{M}^d, \mathbf{M}^p \rangle$ .

A1 does not receive supervised data regarding the words that are appropriate in pictures, nor regarding which words should be used to refer to the referent. The only supervision regarding the “goodness” of words to the given  $\langle \text{referent}, \text{context} \rangle$  pair comes from the success of the interaction between the agents while playing the referential game.

**Agent A2 (Reference Resolution)** For the purposes of the game, A2 needs to perform a task similar to reference resolution. Given the same visual input as A1 (we assume that the agents share the same visual system) and the produced word  $a$ , A2 has to choose which of the 2 objects in the scene is the intended referent. Note that A2 sees the two images in random order.

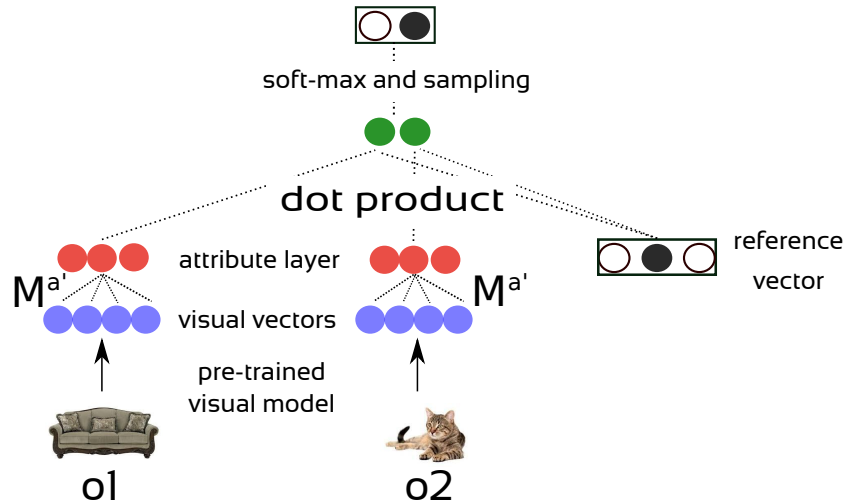


Figure 5-3: Network of A2. In this example, given the two objects and the reference vector encoding the predicted word produced by A1, she correctly predicts that the referent is the second object.

Following this reasoning, we design a simple implementation of A2 depicted in Figure 5-3. A2 is presented with the two objects  $o_1$  and  $o_2$ , without knowing which is

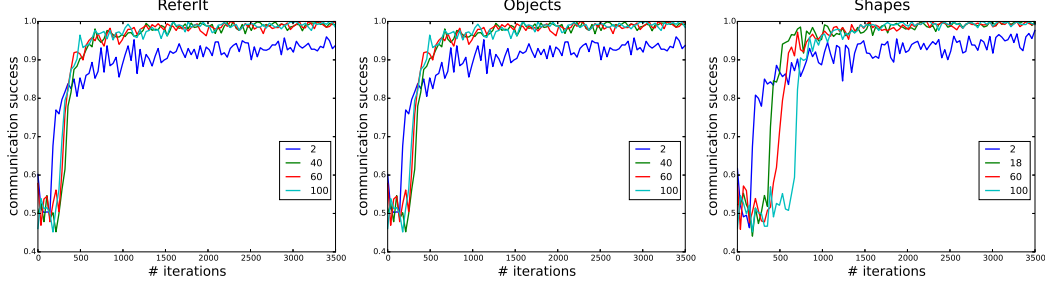


Figure 5-4: Communication success across different datasets over training epochs. For each dataset, we vary the number of words available in the vocabulary of the agents, with the corresponding curves depicted in different colors in the figure *Note: Best viewed in color.*

the referent, and embeds them into a word space using weights shared across the two objects  $\mathbf{M}^{\mathbf{a}'} \in \mathbf{R}^{4096 \times |V|}$ .<sup>3</sup> Note, that as is in the case of A1, A2 will receive no direct image-word supervision. The resulting word vectors encode how active the words are across the two objects. Following that, A2 computes the dot product similarities between the reference vector (i.e., the one-hot representation of the selected word  $a$ ) and the word vectors of the objects. Intuitively, the reference vector encodes which word characterizes the referent best in the current context and as a result, the dot similarity will be high if the word  $a$  is very active in the word vector. These two dot similarities are converted to a probability distribution  $p(o|o_1, o_2, a)$  over the two image indices, and one index is sampled indicating which of the two objects is the chosen referent. The learnable parameters of A2 are just  $\theta_{A2} = \langle \mathbf{M}^{\mathbf{a}'} \rangle$

## 5.3 Experiments

### 5.3.1 General Training Details

The parameters of the 2 agents,  $\theta = \langle \theta_{A1}, \theta_{A2} \rangle$ , are learned jointly while playing the game. The only supervision used is communication success, i.e., whether the agents agreed on the referent. This setup can be naturally modeled with Reinforcement

<sup>3</sup>While we could tie  $\mathbf{M}^{\mathbf{a}'}$  and  $\mathbf{M}^{\mathbf{a}}$ , here we do not enforce any such constraint, essentially allowing the agents to develop their own “visual” understanding.

Learning [102]. Under this framework, the parameters of the agents implement a *policy*. By executing this policy, the agents perform *actions*, i.e., A1 picks a word and A2 picks an image. The loss function that the two agents are minimizing is  $-\mathbf{E}_{\tilde{o} \sim p(o|o_1, o_2, a)}[R(\tilde{o})]$ , where  $o_1$  and  $o_2$  are the 2 objects in the visual scene,  $p(o|o_1, o_2, a)$  is the conditional probability over the 2 objects as computed by A2 given the objects and the word produced by A1, and  $R$  is the reward function which returns 1 iff  $\tilde{o} =$  referent. The parameter updates are done following the Reinforce update rule [114]. We do mini-batch updates, with a batch-size of 32 and train in all datasets for 3.5k iterations.<sup>4</sup>

The agents are trained and tested separately within each dataset. At test time, visual scenes (i.e., combination of referent and contexts images) are novel but individual images might be familiar. For test and tuning we use 1k visual scenes, and leave the rest for training (see Table 5.1 for exact numbers).

### 5.3.2 Results

Our pilot experiments aim to ascertain whether our proposal can result in agents that learn to play the game correctly. Moreover, given that the agents start from a clean state (i.e., they possess no prior semantics other than the relatively low-level features passed to them through the visual vectors), it is worth looking into the nature of the induced semantics of the words. Simply put, is communication-based learning enough to allow agents to use the words in such a way that reflects high-level understanding of the images?

**Can agents learn to develop a communication protocol within the referential game?** Figure 5-4 shows the communication success performance, i.e., how often the intended referent was guessed correctly by A2 (chance guessing would lead to 0.5 performance). Overall, agents are able to come up with a communication protocol allowing them to solve the task, but it takes them approximately 500 iterations (i.e., 16k training examples) before they start communicating effectively. This is to

---

<sup>4</sup>The model-specific hidden size of discriminative layer hyperparameter is set to 20.

be expected, as the agents have no knowledge of the game rules nor of how to refer to things, and must induce both by observing the rewards they receive. Moreover, fewer words in the vocabulary translates to faster, but not necessarily better learning. As a sanity check, we restricted the vocabulary to 2 words only. In this case, the agents also came close to solving the task (and that was done faster than in the other cases), although without approaching 100% performance, a tendency observed across the datasets. At first glance this might seem suspicious; even if we, as humans, use language flexibly through polysemy, still it will not be possible to come up with 2 words being able to reliably distinguish all possible combinations of objects!

However, when we closely inspected the way A1 used the 2 words (e.g., by looking into the induced weights  $\mathbf{M}^a$ ), it became clear that the agent was in a sense “cheating”. Instead of communicating about high-level semantics, the agents agreed to exploit the words to communicate about low-level embedding properties of  $\langle \text{referent, context} \rangle$  pairs (e.g., pick word 1 if the value in dimension 3 is greater in the referent than the context etc). While this might seem odd, such strategies are in fact the best in order to communicate efficiently with only 2 words. This resembles the so-called *conceptual pacts* that humans form to make conversation more efficient, i.e., mutually agreeing in using “unconventional” semantics to refer to things [14]. Still, the “words” discovered in this way have a very ad-hoc meaning that will not generalize to any useful task beyond the specifics of our game, and we would not want the agents to learn them. We thus turn now to an analysis of the semantic nature of the induced words when the agents have a larger vocabulary available than just 2 words, to see if they learn more general meanings, corresponding to (clusters of) high-level visual properties such as “red” or “cat”.

**What is the meaning of the induced words?** Revealing the semantics assigned to the induced words  $a$  is not trivial. We tested whether the semantics of the induced words align with the semantics of the referring expressions as expressed by the gold terms. We focused on the **Objects** and **Shapes** datasets that have a relatively small number of words (100 and 18 respectively) and trained the agents using as word

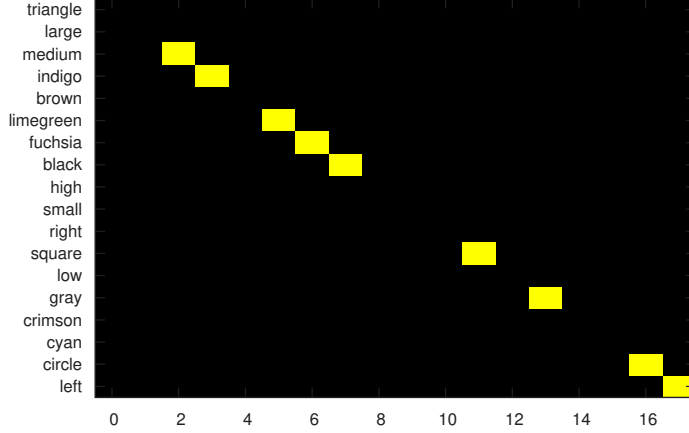


Figure 5-5: Inferred alignment of gold terms to the induced ones in the **Shapes dataset**.

vocabulary size  $|V|$  100 and 18 respectively. After the training, we assign to each induced word  $a$  the gold term that appears most often in the annotations of the  $\langle \text{referent}, \text{context} \rangle$  pairs for which  $a$  was predicted by A1, enforcing a 1-1 mapping (an induced word can be paired only with one gold word). A1 had a tendency not to make use of all the available words  $a$  in the vocabulary, thus either using words in a polysemous way, or assigning them some semantics different than the one that the gold terms encode. As an example, we plot in Figure 5-5 the inferred alignment between induced and gold terms in the **Shapes** dataset.

Irrespective of the exact interpretation of words, meanings induced within this referential framework should be consistent across  $\langle \text{referent}, \text{context} \rangle$  pairs. For example, if an agent used the word *red* to refer to the object X in the context of Y, then *red* cannot be used again to refer to the object Z in the context of X, since *red* is also a property of the latter. However, the “cheating” approach we reported above for the 2-word case, in which the same word is used communicate about whether a low-level feature is higher in the referent or the context, would not respect this consistency constraint, as X might have a higher value of the relevant feature when compared to Y but lower with respect to Z.

To capture this, we devised a measure that we termed “referential inconsistency” ( $RI$ ). Specifically, for each image  $i$  we compute a set  $R(i)$  containing all the induced

<div>datasets</div> <div>words</div>	ReferIt	Objects	Shapes
2	1.0	1.0	1.0
100	0.0	0.03	0.05

Table 5.2: Proportion of referentially inconsistent words, i.e.,  $RI(a) > 0$ , across different datasets and with different word vocabulary size  $|V|$ . Smaller values reflect consistent use.

words that were activated when  $i$  was in the referent slot and  $C(i)$  in the context slot. Then, the referential inconsistency  $RI$  of a word  $a$  is computed as  $\frac{\sum_i [a \in R(i) \cap C(i)]}{\sum_i [a \in R(i) \cup C(i)]}$ , which counts the number of images for which  $a$  was both in the referent and context sets, normalized by the times it appears in either. Ideally, this value should be 0, as this happens when it was never the case that an induced word was activated both when an image was in the referent and in the context slot.

Table 5.2 reports the proportion of induced active words that have  $RI > 0$  across datasets (smaller values reflect consistent use of words). As expected, when communicating with 2 words only, agents do not seem to be using them in a semantically meaningful way, since referential consistency is violated. However, we ought to mention that it is also possible is that the agents in this case have learned to assign a comparative meaning to words [76]. Imagine that we pair an image of Mona Lisa once with an image of a frowning face as the context, and once with an image of a broadly smiling face. It would be acceptable then to use “the smiling one” to denote Mona Lisa in the first pair, but it would also be possible to refer in this way to the more overtly smiling face in the second case, as Mona Lisa is more smiley than the frowning face, but definitely less so than a fully smiling face! In any case, for all datasets, with 100 words referential consistency is largely respected.

Finally, we consider a third way to assess the degree to which the induced words reflect the intuitive semantic properties of the images. Our hypothesis is that, if the induced words are used coherently across visually similar referents, then they should reflect properties that are typical of the class shared by the referents (e.g., “furry” for mammals). For this experiment, we focus on the **Objects** dataset that is annotated

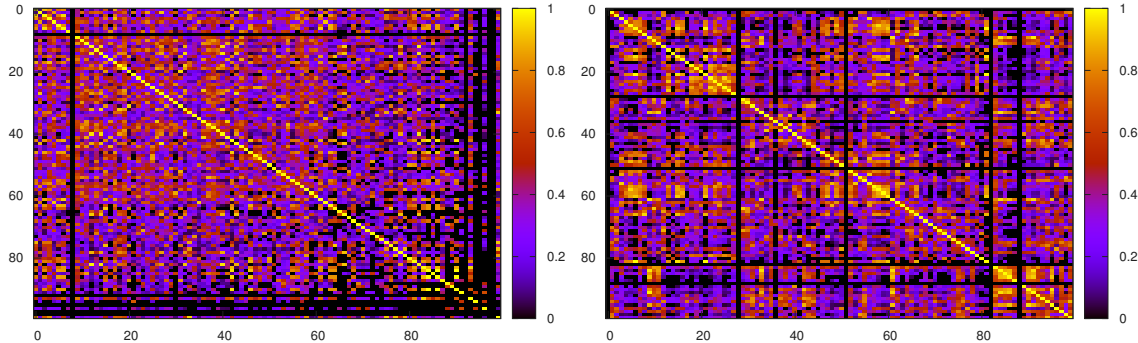


Figure 5-6: Pairwise cosine similarities of gold terms in induced-word vector space in the **Objects** dataset when ordered randomly (**up**) and according to their category (**down**).

with 100 gold terms denoting the objects depicted in the pictures (e.g., *cat*, *dog* etc). For each gold term  $g$ , we construct a vector that records how often the induced word  $a$  was used for a  $\langle \text{referent}, \text{context} \rangle$  pair that was annotated with  $g$ , essentially representing gold terms in a vector space with induced words as dimensions. We then compute the pairwise cosine similarities of the gold terms in this vector space, plotted in Figure 5-6 **up**. As is, there is no structure in the similarity matrix. However, if we organize the rows and columns in the similarity matrix, as in Figure 5-6 **down**, so that objects of the same category cluster together (e.g., the first 2 rows and columns correspond to *appliances*, the next 4 to *fruits*), then a pattern along the diagonal starts to emerge, suggesting that the induced words reflect, at least to some degree, the similarity that exists between objects of the same category.

## 5.4 Discussion

We have presented here a proposal for developing intelligent agents with language capabilities, that breaks away from current passive supervised regimes. Agents co-exist and are able to interact with each other. In our proposed framework we do not restrict the number of agents, nor their role in the games, i.e., we envision a *community* of agents that all interact with each other having to perform different tasks and taking turns in them, requiring them to either co-operate or antagonize in *min-max* sort of *zero-sum* games, in which agents aim at minimizing the opponents

gain (e.g., as in the case of the famous *tic-tac-toe* game).

In our test case, we considered the most basic act of communication, i.e., learning to refer to things, and we designed a “grounded” co-operative task that takes the form of referential games played by two agents. The first experiments, while encouraging, have revealed that it is essential to ensure that agents will not “drift” into their own language, but instead they will evolve one that is aligned to our natural languages. Thus, inspired by the success of AlphaGo [91], that combines both passive learning (experiencing past human games and using a convolutional network to learn valid moves) and interactive learning (learning by playing what is the best move given a particular situation/state), we believe that it is crucial to combine dynamic interactive learning with static statistical learning of patterns from association, something that should ensure the grounding of communication into natural language.

We plan to move along a similar direction, introducing our agents to multi-tasking. As an example, we could expose A1 to large collections of texts and train her on language modeling, a task requiring no manual annotation, from which basic word associations patterns can be learned. Similarly, A2 could be trained on an image retrieval task, from which basic concept recognition and naming capabilities can be acquired, i.e., associating the phonetic string “cat” to instances of cats. Still, the agents would be trained via playing the game for producing good referring expressions, which is a task that depends predominantly on the success of communication (i.e., did our listener understood what we were referring to?)



# Chapter 6

## Conclusion

In this work, we have presented three models of human learning from naturalistic multi-modal input. We first introduced the MULTIMODAL SKIP-GRAM MODEL (Chapter 2), a model that assumes a purely predictive learner existing in a non-communicative setup and showed that such a computational learner displays comparable learning behaviour as human learners on a novel word learning setup (cf. Chapter 3). In Chapter 4, we relaxed some of the learning assumptions and presented the ATTENTIVE SOCIAL MSG, which instead of exposing the computational learner to a passive environment, such as text corpora traditionally used in semantic learning experiments, it exposes the learner in communicative episodes, simulated in our experiments by corpora capturing multi-modal interactions between children and their caregivers, allowing the learner to make use of information beyond words and passive percepts during learning. Finally, in Chapter 5, we presented on-going work towards fully interactive learning between two agents who learned to develop a language through the need to communicate in order to co-operatively solve a task.

We believe that the most promising direction towards developing agents that can work together with us is to place emphasis on the interactive aspect of language. After all, human-machine co-ordination, just like the human-human one, is impossible without communication, and the bulk of communication happens through natural language, supplemented of course with other non-linguistic cues contributing to rich multi-modal interactions. Along these lines, we advocate a research program that

creates interactive environments between a community of artificial agents. The agents will be given tasks to solve (i.e., games to play similar to the ones we presented in Chapter 5), that will be designed in order to foster emergence of key properties of language. Language will not be the goal of these games, but rather the by-product of achieving situated goals in this environment.

Connecting the emerged communication to human natural language is a key challenge for this framework; supervised learning and interactive learning should both come together to achieve this goal. We think that predictive learning should be retained as an important building block of intelligent agents, but rather should be focused on teaching them primitive properties of language (e.g., basic word-object associations as we also did in Chapters 2 and 4) rather than complex behaviours, such as how to hold a conversation [108].

# Bibliography

- [1] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Learning to compose neural networks for question answering. *arXiv preprint arXiv:1601.01705*, 2016.
- [2] Mark Andrews, Gabriella Vigliocco, and David Vinson. Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, 116(3):463–498, 2009.
- [3] John Langshaw Austin. *How to do things with words*. Harvard University Press, Cambridge, MA, 1962.
- [4] R. H. Baayen, D. J. Davidson, and D.M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, pages 390–412, 2008.
- [5] R Harald Baayen, Petar Milin, Dusica Filipović Durdević, Peter Hendrix, and Marco Marelli. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*, 118(3):438, 2011.
- [6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR Conference Track*, San Diego, CA, 2015. Published online: <http://www.iclr.cc/doku.php?id=iclr2015:main>.
- [7] Marco Baroni, Eduard Barbu, Brian Murphy, and Massimo Poesio. Strudel: A distributional semantic model based on properties and types. *Cognitive Science*, 34(2):222–254, 2010.
- [8] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don’t count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, pages 238–247, Baltimore, MD, 2014.
- [9] Marco Baroni, Alessandro Lenci, and Luca Onnis. ISA meets Lara: An incremental word space model for cognitively plausible simulations of semantic learning. In *Proceedings of the ACL Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 49–56, 2007.

- [10] Lawrence Barsalou. Grounded cognition. *Annual Review of Psychology*, 59:617–645, 2008.
- [11] Lawrence Barsalou and Katja Wiemer-Hastings. Situating abstract concepts. In D. Pecher and R. Zwaan, editors, *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thought*, pages 129–163. Cambridge University Press, Cambridge, UK, 2005.
- [12] Arielle Borovsky, Jeffrey Elman, and Marta Kutas. Once is enough: N400 indexes semantic integration of novel word meanings from a single exposure in context. *Language Learning and Development*, 8(3):278–302, 2012.
- [13] Arielle Borovsky, Marta Kutas, and Jeff Elman. Learning to use words: Event related potentials index single-shot contextual word learning. *Cognition*, 116(2):289–296, 2010.
- [14] Susan E Brennan and Herbert H Clark. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482, 1996.
- [15] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. Distributional semantics in Technicolor. In *Proceedings of ACL*, pages 136–145, Jeju Island, Korea, 2012.
- [16] Elia Bruni, Nam Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47, 2014.
- [17] Murray Campbell, A Joseph Hoane, and Feng-hsiung Hsu. Deep blue. *Artificial intelligence*, 134(1):57–83, 2002.
- [18] Malinda Carpenter, Katherine Nagell, Michael Tomasello, George Butterworth, and Chris Moore. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, 63(4), 1998.
- [19] Stephen Clark. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantics*, 2nd ed., pages 493–522. Blackwell, Malden, MA, 2015.
- [20] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of ICML*, pages 160–167, Helsinki, Finland, 2008.
- [21] Robert Dale and Nicholas Haddock. Content determination in the generation of referring expressions. *Computational Intelligence*, 7(4):252–265, 1991.
- [22] Robert Dale and Ehud Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263, 1995.

- [23] Jia Deng, Wei Dong, Richard Socher, Lia-Ji Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of CVPR*, pages 248–255, Miami Beach, FL, 2009.
- [24] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159, 2011.
- [25] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.
- [26] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Proceedings of CVPR*, pages 1778–1785, Miami Beach, FL, 2009.
- [27] Afsaneh Fazly, Afra Alishahi, and Suzanne Stevenson. A probabilistic computational model of cross-situational word learning. *Cognitive Science*, 34:1017–1063, 2010.
- [28] Yansong Feng and Mirella Lapata. Visual information in semantic representation. In *Proceedings of HLT-NAACL*, pages 91–99, Los Angeles, CA, 2010.
- [29] Michael Frank, Noah Goodman, and Joshua Tenenbaum. A Bayesian framework for cross-situational word-learning. In *Proceedings of NIPS*, pages 457–464, Vancouver, Canada, 2007.
- [30] Diego Frassinelli and Frank Keller. The plausibility of semantic properties generated by a distributional model: Evidence from a visual world experiment. In *Proceedings of CogSci*, pages 1560–1565, 2012.
- [31] Andrea Frome, Greg Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc’Aurelio Ranzato, and Tomas Mikolov. DeViSE: A deep visual-semantic embedding model. In *Proceedings of NIPS*, pages 2121–2129, Lake Tahoe, NV, 2013.
- [32] Alona Fyshe, Partha P Talukdar, Brian Murphy, and Tom M Mitchell. Interpretable semantic vectors from a joint model of brain-and text-based meaning. In *In Proceedings of ACL*, pages 489–499, 2014.
- [33] Bruno Galantucci, Simon Garrod, and Gareth Roberts. Experimental semiotics. *Language and Linguistics Compass*, 6(8):477–493, 2012.
- [34] Jane Gillette, Henry Gleitman, Lila Gleitman, and Anne Lederer. Human simulations of vocabulary learning. *Cognition*, 73(2):135–176, 1999.
- [35] Arthur Glenberg and David Robertson. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, 3(43):379–401, 2000.

- [36] Susan Goldin-Meadow. *Hearing gesture: How our hands help us think*. Harvard University Press, 2005.
- [37] Tom Griffiths, Mark Steyvers, and Josh Tenenbaum. Topics in semantic representation. *Psychological Review*, 114:211–244, 2007.
- [38] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.
- [39] David A Havas, Arthur M Glenberg, and Mike Rinck. Emotion simulation during language comprehension. *Psychonomic Bulletin & Review*, 14(3):436–441, 2007.
- [40] Felix Hill, KyungHyun Cho, Sébastien Jean, Coline Devin, and Yoshua Bengio. Not all neural embeddings are born equal. In *Proceedings of the NIPS Learning Semantics Workshop*, Montreal, Canada, 2014. Published online: <https://sites.google.com/site/learningsemantics2014/>.
- [41] Felix Hill and Anna Korhonen. Learning abstract concept embeddings from multi-modal data: Since you probably can’t see what I mean. In *Proceedings of EMNLP*, pages 255–265, Doha, Qatar, 2014.
- [42] Steve Howell, Damian Jankowicz, and Suzanna Becker. A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *Journal of Memory and Language*, 53:258–276, 2005.
- [43] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.
- [44] Ákos Kádár, Afra Alishahi, and Grzegorz Chrupała. Learning word meanings from images of natural scenes. *Traitement Automatique des Langues*, 2015. In press, preprint available at <http://grzegorz.chrupala.me/papers/tal-2015.pdf>.
- [45] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Proceedings of NIPS*, pages 1097–1105, Montreal, Canada, 2014.
- [46] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. Refer-itgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014.
- [47] Casey Kennington and David Schlangen. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the Conference for the Association for Computational Linguistics (ACL)*, 2015.

- [48] Emmanuel Keuleers and Marc Brysbaert. Wuggy: A multilingual pseudoword generator. *Behavior research methods*, 42(3):627–633, 2010.
- [49] Douwe Kiela and Léon Bottou. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of EMNLP*, pages 36–45, Doha, Qatar, 2014.
- [50] Douwe Kiela, Felix Hill, Anna Korhonen, and Stephen Clark. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of ACL*, pages 835–841, Baltimore, MD, 2014.
- [51] Brent Kievit-Kylar and Michael Jones. The Semantic Pictionary project. In *Proceedings of CogSci*, pages 2229–2234, Austin, TX, 2011.
- [52] Brent Kievit-Kylar, George Kachergis, and Michael Jones. Naturalistic word-concept pair learning with semantic spaces. In *Proceedings of CogSci*, pages 2716–2721, Berlin, Germany, 2013.
- [53] Ryan Kiros, Ruslan Salakhutdinov, and Richard Zemel. Unifying visual-semantic embeddings with multimodal neural language models. In *Proceedings of the NIPS Deep Learning and Representation Learning Workshop*, Montreal, Canada, 2014. Published online: <http://www.dlworkshop.org/accepted-papers>.
- [54] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of NIPS*, pages 1097–1105, Lake Tahoe, Nevada, 2012.
- [55] Brenden Lake, Tomer Ullman, Joshua Tenenbaum, and Samuel Gershman. Building machines that learn and think like people. <https://arxiv.org/abs/1604.00289>, 2016.
- [56] George Lakoff and Mark Johnson. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. Basic Books, New York, 1999.
- [57] Thomas Landauer and Susan Dumais. A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240, 1997.
- [58] Angeliki Lazaridou, Elia Bruni, and Marco Baroni. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of ACL*, pages 1403–1414, Baltimore, MD, 2014.
- [59] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Combining language and vision with a multimodal skip-gram model. In *Proceedings of NAACL*, pages 153–163, Denver, CO, 2015.

- [60] Angeliki Lazaridou, Dat Tien Nguyen, and Marco Baroni. Do distributed semantic models dream of electric sheep? Visualizing word representations through image synthesis. In *Proceedings of the EMNLP Vision and Language Workshop*, pages 81–86, Lisbon, Portugal, 2015.
- [61] Alessandro Lenci. Distributional approaches in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1):1–31, 2008.
- [62] Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, 28:203–208, 1996.
- [63] Brian MacWhinney. *The CHILDES Project: Tools for analyzing talk*. Lawrence Erlbaum Associates, 3rd edition, 2000.
- [64] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan Yuille. Explain images with multimodal recurrent neural networks. In *Proceedings of the NIPS Deep Learning and Representation Learning Workshop*, Montreal, Canada, 2014. Published online: <http://www.dlworkshop.org/accepted-papers>.
- [65] Scott McDonald and Michael Ramscar. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of CogSci*, pages 611–616, 2001.
- [66] Ken McRae, George Cree, Mark Seidenberg, and Chris McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559, 2005.
- [67] Anna Mestres-Missé, Antoni Rodriguez-Fornells, and Thomas Münte. Watching the brain during meaning acquisition. *Cerebral Cortex*, 17(8):1858–1866, 2007.
- [68] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. <http://arxiv.org/abs/1301.3781/>, 2013.
- [69] Tomas Mikolov, Armand Joulin, and Marco Baroni. A roadmap towards machine intelligence. *arXiv preprint arXiv:1511.08130*, 2015.
- [70] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*, pages 3111–3119, Lake Tahoe, NV, 2013.
- [71] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL*, pages 746–751, Atlanta, Georgia, 2013.
- [72] Margaret Mitchell, Kees van Deemter, and Ehud Reiter. Natural reference to objects in a visual domain. In *Proceedings of the 6th international natural language generation conference*, pages 95–104. Association for Computational Linguistics, 2010.



- [73] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei Rusu, Joel Veness, Marc Bellemare, Alex Graves, Martin Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharmashan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.
- [74] Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In *Proceedings of AISTATS*, pages 246–252, Barbados, 2005.
- [75] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of CVPR*, 2014.
- [76] Devi Parikh and Kristen Grauman. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 503–510. IEEE, 2011.
- [77] Mary C Potter, Judith F Kroll, Betsy Yachzel, Elisabeth Carpenter, and Janet Sherman. Pictures in sentences: Understanding without words. *Journal of Experimental Psychology: General*, 115(3):281, 1986.
- [78] Friedemann Pulvermüller, Olaf Hauk, Vadim V Nikulin, and Risto J Ilmoniemi. Functional links between motor and language systems. *European Journal of Neuroscience*, 21(3):793–797, 2005.
- [79] Adria Recasens\*, Aditya Khosla\*, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems (NIPS)*, 2015. \* indicates equal contribution.
- [80] Leslie Rescorla. Overextension in early language development. *Journal of Child Language*, 7(2):321–335, 1980.
- [81] Brandon C. Roy, Michael C. Frank, Philip DeCamp, Matthew Miller, and Deb Roy. Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, 112(41):12663–12668, 2015.
- [82] Deb Roy. A computational model of word learning from multimodal sensory input. In *Proceedings of the International Conference of Cognitive Modeling (ICCM2000)*, Groningen, Netherlands, 2000.
- [83] Deb Roy. New horizons in the study of child language acquisition. In *Proceedings of Interspeech*, 2009.
- [84] Margarita Sarri, Richard Greenwood, Lalit Kalra, and Jon Driver. Prism adaptation does not change the rightward spatial preference bias found with ambiguous stimuli in unilateral neglect. *Cortex*, 47(3):353–366, 2011.
- [85] Tyler Schnoebelen and Victor Kuperman. Using Amazon Mechanical Turk for linguistic research. *Psihologija*, 43(4):441–464, 2010.

- [86] Hinrich Schütze. *Ambiguity Resolution in Natural Language Learning*. CSLI, Stanford, CA, 1997.
- [87] John Searle. *Minds, Brains and Science*. Harvard University Press, Cambridge, MA, 1984.
- [88] Carina Silberer, Vittorio Ferrari, and Mirella Lapata. Models of semantic representation with visual attributes. In *Proceedings of ACL*, pages 572–582, Sofia, Bulgaria, 2013.
- [89] Carina Silberer and Mirella Lapata. Grounded models of semantic representation. In *Proceedings of EMNLP*, pages 1423–1433, Jeju, Korea, 2012.
- [90] Carina Silberer and Mirella Lapata. Learning grounded meaning representations with autoencoders. In *Proceedings of ACL*, pages 721–732, Baltimore, Maryland, 2014.
- [91] David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503, 2016.
- [92] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [93] Josef Sivic and Andrew Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of ICCV*, pages 1470–1477, Nice, France, 2003.
- [94] Brian Skyrms. *Signals: Evolution, learning, and information*. Oxford University Press, 2010.
- [95] Linda Smith, Sumarga Suanda, and Chen Yu. The unrealized promise of infant statistical word-referent learning. *Trends in Cognitive Sciences*, 18(5):251–258, 2014.
- [96] Richard Socher, Danqi Chen, Christopher Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In *Proceedings of NIPS*, pages 926–934, Lake Tahoe, NV, 2013.
- [97] Richard Socher, Milind Ganjoo, Christopher Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Proceedings of NIPS*, pages 935–943, Lake Tahoe, NV, 2013.
- [98] Richard Socher, Quoc Le, Christopher Manning, and Andrew Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

- [99] Luc Steels. *The Talking Heads experiment: Origins of words and meanings*, volume 1. Language Science Press, 2015.
- [100] Tanya Stivers and Jack Sidnell. Introduction: Multimodal interaction. *Semiotica*, pages 1–20, 2005.
- [101] Ilya Sutskever, Oriol Vinyals, and Quoc Le. Sequence to sequence learning with neural networks. In *Proceedings of NIPS*, pages 3104–3112, Montreal, Canada, 2014.
- [102] Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.
- [103] Stefanie Tellex, Ross Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. Asking for help using inverse semantics. 2014.
- [104] Gerald Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3):58–68, 1995.
- [105] John Trueswell, Tamara Medina, Alon Hafri, and Lila Gleitman. Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, 66(1):126–156, 2013.
- [106] Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of EMNLP*, pages 680–690, Edinburgh, UK., 2011.
- [107] Peter Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.
- [108] Oriol Vinyals and Quoc Le. A neural conversational model. In *Proceedings of the ICML Deep Learning Workshop*, Lille, France, 2015. Published online: <https://sites.google.com/site/deeplearning2015/accepted-papers>.
- [109] Haley Vlach and Catherine Sandhofer. Developmental differences in children’s context-dependent word learning. *Journal of Experimental Child Psychology*, 108(2):394–401, 2011.
- [110] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of CHI*, pages 319–326, Vienna, Austria, 2004.
- [111] Kyle Wagner, James A Reggia, Juan Uriagereka, and Gerald S Wilkinson. Progress in the simulation of emergent communication and language. *Adaptive Behavior*, 11(1):37–69, 2003.
- [112] Jason Weston. Dialog-based language learning. *arXiv preprint arXiv:1604.06045*, 2016.

- [113] Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010.
- [114] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.
- [115] Terry Winograd. Procedures as a representation for data in a computer program for understanding natural language. Technical Report AI 235, Massachusetts Institute of Technology, 1971.
- [116] Ludwig Wittgenstein. *Philosophical Investigations*. Blackwell, Oxford, UK, 1953. Translated by G.E.M. Anscombe.
- [117] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of ICML*, pages 2048–2057, Lille, France, 2015.
- [118] C. Yu. The emergence of links between lexical acquisition and object categorization: A computational study. *Connection Science*, 17(3):381–397, 2005.
- [119] Chen Yu and Dana H. Ballard. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15):2149–2165, 2007.
- [120] Chen Yu and Linda B. Smith. Joint attention without gaze following: human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PLoS ONE*, 8(11), 2013.
- [121] Matthew Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of ECCV (Part 1)*, pages 818–833, Zurich, Switzerland, 2014.